André Sandberg

Academic Vocabulary in Argumentative Essays of Finland-Swedish Intermediate Learners

André Sandberg
Pro gradu-avhandling i engelska språket och litteraturen
Handledare: Brita Wårvik
Fakulteten för humaniora, psykologi och teologi
Åbo Akademi
2018

André Sandberg

ÅBO AKADEMI UNIVERSITY – THE FACULTY OF ARTS, PSYCHOLOGY AND THEOLOGY

Subject: English Language and Literature
Author: André Sandberg
Title: Academic Vocabulary in Finland-Swedish Learners' Essays
Supervisor: Brita Wårvik
The purpose of the present study is to examine the academic vocabulary in Finland-Swedish learners' argumentative essays. Requirements on students' abilities to write and speak in English increase day by day. The ability to write correctly and to use the proper vocabulary implies that students, and also language learners, must master the writing conventions of several genres. Academic writing, including the vocabulary that is associated with it, remains one of the most important genres to learn due to it being applicable in contexts outside of academia. Argumentative writing on the other hand shares many elements with academic writing, and it is a crucial element of EFL learning in upper secondary schools in Finland. A learner corpus, the F-SCUSSE (*Finland-Swedish Corpus of Upper Secondary School Essays*), was compiled for the present study and it consists at the moment of 239 argumentative essays totalling approximately 63,000 words. The EFL learners represent two levels and the learners that constitute those levels originate from several upper secondary schools in the Swedish-speaking parts of Finland. The implementation of two academic vocabulary lists, the AWL (Coxhead, 2000) and the AVL (Gardner & Davies, 2013), make possible comparisons of academic vocabulary use across two levels of non-native speaker learners. The LOCNESS (*Louvain Corpus of Native English Essays*) A-levels essays and the BNC (*British National Corpus*) are used as reference corpora. The software that is utilized in the process of data collection is WordSmith Tools (Scott, 2018). The results show that Finland-Swedish learners of English use a more varied language than what could have been expected, but some language features distinguish them from the English students. Such examples include a higher use of personal pronouns, but a lower use of the *of* genitive case than English students. In addition, there are traces of *lexical teddy bears* (Hasselgren, 1994), although they also occur in essays written by students whose first language is English. The results also suggest that the use of academic vocabulary depends on the variables of at least native speaker/non-native speaker status and the level of the essay writers. Furthermore, it seems that native speaker learners use a more varied and lexically complex language than non-native speaker learners. However, the results show that especially the higher level Finland-Swedish learners and the English students share a number of characteristics in their use of argumentative English in writing. Thus, this research provides valuable information and hints that vocabulary knowledge could relate to reasons other than merely students' first language.
Key words: corpus linguistics, learner corpus research, EFL, F-SCUSSE, academic vocabulary, AWL, AVL, LOCNESS, BNC, word frequencies, WordSmith Tools
Date: 14.12.2018 | Pages: 92

Table of Contents

List of Abbreviations

AVL	Academic Vocabulary List
AWL	Academic Word List
BNC	British National Corpus
EFL	English as a Foreign Language
EAP	English for Academic Purposes
EGAP	English for General Academic Purposes
F-SCUSSE	Finland-Swedish Corpus of Upper Secondary School English
ICLE	International Corpus of Learner English
L1	First Language
L2	Second Language
LOCNESS	Louvain Corpus of Native English Essays
LCR	Learner Corpus Research
NS	Native Speaker (of English)
NNS	Non-Native Speaker (of English)
SLA	Second Language Acquisition
UWL	University Word List

List of Tables and Figures

Tables:

Figures:

1. Introduction

The requirements on students' abilities to write and speak in English increase day by day. The ability to write correctly and to use the proper vocabulary means that students, as well as language learners, must master the writing conventions of several genres. Academic writing, including the vocabulary that is associated with it, remains one of the most important genres to learn due to it being applicable in contexts outside of academia. Assuming that there are roughly a quarter of a million words in the English language (*OED*, 2017), it is definitely relevant that only a small percentage of those words covers a huge majority of the words we encounter on a daily basis. This is subsequently why corpus linguistics, and its sub-field *learner corpus research*, LCR, have become quite influential. Furthermore, the most commonly used words might be a primary goal for the foreign language classroom, as the learner is most likely to encounter such words in his or her daily life. However, where to do teachers and educators turn after they establish that small percentage of words in the classroom? Academic vocabulary might well be a reasonable option.

The present study utilizes corpora to study the language of intermediate EFL learners. The use of corpora, as they are used in corpus linguistics and its sub-fields, is not new. According to Scott and Tribble (2006: 4), the collection and study of large amounts of text was recognized as a great source of knowledge and teaching of, for instance rhetorical methods, many centuries before electronic texts. Leech (2011: 8) describes the approaches of those working with large text collections prior to the computer-age as follows:

> Before computers when researchers counted frequencies by hand, the simple postulate justifying this effort was: 'more frequent = more important to learn'.

Interestingly, even before the introduction of computers, which automated the process of word calculation, it seems that the most common or the highest frequent items were of the biggest interest. It could however be argued that more frequent items are not just more important to learn but also easier to memorize, as they will be encountered more often. Nowadays, the most frequent items are of great value in LCR, although simultaneously, the object of study might not be very frequent. Similarly, words that

are not very frequent may be relevant to learn for some learners, depending on their life situation or personal interests.

As time has passed, we now have large corpora available at everyone's disposal, such as the *British National Corpus*, BNC. The BNC is a practical collection of texts for looking up actual uses in the British variety of English, as it consists of nearly 100 million words. It consists primarily of written language but also of spoken data converted to text. For instance, whenever we encounter a word that may seem outdated or a phrase that seems unconventional, it may be worth to check the items' occurrences in the BNC. The large corpus is therefore used, for example, as a tool when creating any materials that relates to language, such as dictionaries and teaching materials.

The present study is a corpus-driven study on Finland-Swedish upper secondary school students' English vocabulary. The study can be described as learner corpus research in the area of English for general academic purposes, EGAP, or as described by Flowerdew (2015: 469), "argumentative essays written on general topics by intermediate-level students". The aim is two-fold: first, to evaluate EFL learners' general vocabulary, and secondly, to evaluate their academic vocabulary with the help of academic word lists. The latter will follow and be made possible as a result of the former. The *Finland-Swedish Corpus of Upper Secondary School English*, F-SCUSSE, is a learner corpus that has been compiled from Finland-Swedish upper secondary school students' English essays for the purpose of this study. The corpus includes writers from two levels: first year students and third year students. The aim is to cross-reference these materials with two independent academic word lists: the *Academic Word List* or AWL (Coxhead, 2000a) and the *New Academic Vocabulary List* or AVL (Gardner & Davies, 2013a). Furthermore, the A-levels students' essays of the *Louvain Corpus of Native English Essays*, LOCNESS (CECL, n.d.), will be used as a reference corpus, as the authors of the A-levels essays share many characteristics with the authors of the essays included in the F-SCUSSE. Lastly, the BNC will be used as reference, due to its large size, and thus its status as representative of the British variety of the English language.

2. Why Learner Corpus Study?

"Written and spoken data produced by learners has always been a key resource for the study of second language acquisition" (Granger et al., 2015: 1). When we study the vocabulary of learners, we occasionally find ourselves in the situation of questioning measurability. How do we know that elicitation tests designed for the purpose of a research study actually show what learners know and can do? (Olsson & Sylvén, 2017: 127). These tests are usually not part of the ordinary syllabus, and as a result, the students are rarely graded for them. Understandably, we can then ask what should motivate the students to perform well in them. In addition, are the results going to reflect *naturalistic language*, i.e. language not influenced by the researcher? Lozano and Mendikoetxea (2013: 66) suggest that much *second language acquisition*, SLA, research "has traditionally relied on elicited experimental data while disfavouring natural language data", which in turn Granger (2002a: 6) explains is a way for the researcher to control variables. Cook (1986: 13), a front figure for SLA research, comments ironically that "[controlled] data has the advantage that it yields the information we are looking for. It has the disadvantage of artificiality". One solution to this problem could be to study genuine text production that has not been influenced by any third party, where the learners choose their own wording (cf. Granger, 2008: 261). For the present study, essay writing in the classroom represents genuine text production, which is usually part of the syllabus and normally graded. Therefore, the essay writing should demonstrate the learners' utmost ability to write in English. I will return to this matter in section 4 when I discuss the matriculation exams, but if we further consider the element of essay writing as practice for one of the most important exams in the lives of many students, it should mean that there is an additional clear motivator behind the essay writing.

Learner Corpus Research (LCR) has developed substantially over the past two decades, but it can still be considered underdeveloped. In short, there seems to be insufficient amount and variety of data for LCR to have fully developed. Learner corpora are described as "electronic collections of authentic, continuous and contextualised foreign or second language texts produced by learners and assembled according to explicit design criteria" (Granger, 2009: 14). This definition includes several crucial LCR elements, and authenticity is a must. Recently, several English learner corpora have been created, most notably the ICLE (*International Corpus of Learner English*) (Granger, Dagneaux & Meunier, 2002) but also smaller ones, such as

the *EVA Corpus of Norwegian School English* (Hasselgren, n.d.). The use of corpora in research has opened up possibilities previously unavailable in language studies. In fact, Granger (2002a) argues that the combination of computer software tools and learner corpus data has already proven and could continue to prove to be a remarkable method in the discovery of linguistic phenomena in learner language.

However, in learner corpus research, there are challenges involved as we try to collect completely natural data. Essays written by learners in the classroom are not strictly natural, because they are part of a task and based upon the educator's guidelines. In the context of course design or classroom teaching, the essay writing could arguably be the closest we will ever come to the natural production by the students. Furthermore, the same conditions apply to NS learners and their essay writing, where guidelines usually decide the directions of the text production. This ambiguity in terms of natural language is acknowledged by Granger (2015: 1) in her definition of learner corpora as "electronic collections of natural, or *near-natural* data produced by foreign or second language (L2) learners" (emphasis added). These restrictions are not all bad and, for instance, enable the researcher to keep some control over the production (Gilquin, 2015: 10). The more naturalistic the language output, the more challenging it will be to uncover and examine the linguistic phenomenon of interest. Nevertheless, the point of naturalistic data is that the procedure of data collection has not been as restrictive as in elicitation tests.

Corpora and concordancers are tools used primarily for research in corpus linguistics, but some argue that the two could be used outside these contexts. Since the 1990s, researchers have suggested that corpora could be used, for instance, by teachers in the classroom. The suggestion is often two-fold, divided into a direct, or explicit, approach, and an indirect, or implicit, approach. Granger (2009: 20) however, uses another terminology and divides them into learner corpora for delayed pedagogical use (DPU) and learner corpora for immediate pedagogical use (IPU). As the names suggest, the former has to do with implications for future purposes, namely the study of corpora so that any precious findings eventually, or at least hopefully, be implemented in the classroom teaching. For instance, the F-SCUSSE corpus might not have any concrete implications for the learners who have written the texts. Instead, the findings could prove valuable to learners in a similar situation, as the aim is to inform the teachers of any relevant findings. This is the typical approach when we deal with learner corpora, but as previously mentioned, according to Granger (2009: 20) and several others (e.g.

Davies, 2004; Ädel, 2010) the direct approach, or IPU corpora, offers possibilities that could be incorporated in language teaching. In that case, the learners are "given hands-on access to corpora" (Ädel, 2010: 40) and the users are at the same time "producers and users of the corpus data" (Granger, ibid.). Such an example would be Seidlhofer (2002), who compiled a corpus of ELF (*English as a Lingua Franca*) use to be able to categorise the features utilized by both young learners and adults with other L1s than English. Her findings would later help her own classroom teaching and understanding of SLA.

2.1 Limitations of a Corpus Study

Whenever one compares two or more corpora, there are a few issues that need to be raised. According to Callies (2015), there are three methodological challenges in particular involved in the process of working with learner corpora. These include

1. the very definition of a learner corpus, and thus its comparability with others;
2. homogeneity vs. variability in learner corpora;
3. proficiency level as a fuzzy variable in learner corpus compilation and analysis

(Callies, 2015: 25)

These three major challenges require a lot of work, and they can be controlled to some extent by careful corpus compilation and thoughtful corpus design. If, for some reason, the collection of data is not done carefully, it could lead to hidden variables, which would then be capable of disturbing the corpus analysis (Leech, 1998: xix). In addition, there is reason to repeat the differences in data collection between a study in LCR and a traditional SLA study. The aspiration to compile naturalistic data in LCR might also be its biggest flaw, as there is no guarantee that any of the linguistic phenomena, or in this case lexicalities, occurs naturally in the learners' naturalistic language production. This is known generally known as *construct underrepresentation* (Callies, 2015: 42). However, whereas the previous challenges are mostly connected to the corpus compilation, Gries stresses the importance of choosing a solid methodology when the corpus analysis is actually performed. "Nothing in linguistics is truly monocausal […,] which means that LCR should embrace methods that can handle *multi*factoriality more" (2015: 175). What Gries implies here is that too often conclusions have been made based on the learners' L1 in LCR, i.e. first language interference, while other causes

may remain unidentifiable or simply overlooked. A well-constructed methodology is thus essential in the context of LCR. In addition, there is a need to highlight the components Granger (2009: 15) indicates as constituents of LCR: corpus linguistics, foreign language teaching, linguistic theory and second language acquisition. Thus, in a conscientious LCR study it is advantageous to have at least some basic understanding of each one of these fields.

Another issue that is worth highlighting relates to the nature of the corpus itself. For instance, any results from a learner corpus might only be applicable to learners of very similar backgrounds. Thus, possible results from a corpus of Finland-Swedish students, such as is the case with the F-SCUSSE, are only applicable to them, and not Swedish students nor any other groups of learners. Additionally, when two or more corpora are compared, for instance one NNS corpus with two NS corpora as reference corpora, the choice of target norm might be problematic (Altenberg, 2011: xv). The deliberate choice, which is based upon the aim of the study, is most influential and might affect the results to a large extent. Thus, any relevant findings must be presented with all of the necessary applied reference corpora in mind. In the methods chapter, I will explain how I intend to handle these methodological challenges.

2.2 Research Questions

In the present study, my aim is to answer four questions. In general, the questions relate to two major themes, or one major theme and a sub-theme: the distinction between native speaker, NS, writing and non-native speaker, NNS, writing, and the young writers' use of academic vocabulary in two individual corpora. These four questions provide a starting point although there certainly will be further questions in connection to these that arise as the study progresses.

1. What is the coverage of the *Academic Word List*, AWL, and the *Academic Vocabulary List*, AVL, in the *Finland-Swedish Corpus of Upper Secondary School English*, F-SCUSSE, in contrast to the A-levels essays of the *Louvain Corpus of Native English Essays*, LOCNESS?
2. What academic lexical items, as suggested by the AWL and the AVL, are the most frequent in the F-SCUSSE?
3. How does the use of academic vocabulary vary between different levels of learners in the F-SCUSSE?

4. How does NNS learner written language compare to NS learner written language, and to what extent does the level of the NNS learners influence this comparison?

Based on the findings of previous studies, and with the methodological challenges in mind, my hypothesis is that some lexical items, which are typically associated with spoken language, would be used in more easily distinguishable ways in the F-SCUSSE in comparison to NS corpora. However, the expectation is to find that the older learners exhibit a better vocabulary knowledge than the younger learners do, for instance in the use of a more carefully constructed register, as they have had more practice in writing argumentatively. Moreover, it would not be surprising to find that all of these students of approximately similar age, no matter if they are NNS or NS learners, would display a similar use of vocabulary as well, especially if we take their similar life situations into consideration. In addition, it could be expected that a few *lexical teddy bears* (Hasselgren, 1994) in particular would appear. Leech (2011: 14) describes lexical teddy bears as "words with which the learners feel most familiar, most confident and most comfortable". Lexical teddy bears are closely tied to the phenomenon of mixing written and spoken language, and seemingly they can be found in both NS and NNS written and spoken language. However, the lexical teddy bears produced by NNS students could display less diversity than the ones produced by NS students, as the NNS students should have a vocabulary that contains fewer items. The phenomenon of lexical teddy bears will continuously be assessed in association to the academic vocabulary throughout this study, especially in connection with the theme of vagueness.

3. Background and Definitions

The purpose of this chapter is to supply sufficient background for the present study, as well as to position it in a suitable context. Moreover, its function is to offer the reader definitions and perspective. The chapter is structured so that vocabulary research comes first, followed by previous learner corpus research, academic vocabulary and lastly, academic word lists. The section on general vocabulary research, with focus on vocabulary learning, can be perceived as a starting point; a background in itself to the more genre-specific sections on LCR, academic vocabulary and academic word lists. The latter two are closely connected but have each been given their own section. In that process, it will also become clear that different disciplines do not always agree with each other completely. Lastly, the Finnish upper secondary school is introduced, as well as the Matriculation Examinations. The Matriculation Examinations function as the final assessment of upper secondary school in Finland. These two elements are presented because they are closely connected to the materials that constitute the F-SCUSSE. The focus is on the school subject English.

3.1 Vocabulary Research

Before we move into the area of corpus linguistics and academic vocabulary, I want to offer the reader some perspective on traditional vocabulary research in connection with the other topics. Vocabulary research has showed us that learners with similar language backgrounds tend to face similar problems when learning a language. Furthermore, as proposed by Mauranen (2011: 158) a classroom with shared linguistic and cultural assumptions makes for an environment where, for instance, expectations relative to the target language are shared as well. In other words, the learners share a common goal. According to Hasselgård and Johansson (2011: 33), “advanced language learners make mistakes and normally have a limited repertoire compared with native speakers of the target language”. Consequently then, learner corpora should be filled with errors. The statement may seem controversial, and it should not be considered as a generalization, although it is not completely untrue. However, native speakers make mistakes too, which subsequently means that native speaker corpora also would contain errors (Granger, in Viana, 2007: 12).

We must remember that to learn a language from birth, in contrast to being taught it as an older child, or as an adult, makes for very different circumstances. The repertoire of language learners may be limited but at the same time, it may be

highly specialized, due to individual reasons. For instance, some learners may have a highly specialized vocabulary thanks to their personal activities. Nevertheless, Hasselgård and Johansson (2011: 33) continue by listing three reasons why problems so often arise when we try to learn a foreign language. These include features of the target language, first language influence, or simply complications within the learning process. These general problems vary from person to person as well as from language to language.

It could be argued that in terms of vocabulary size, we can consider an expanded vocabulary as equal to a near-native fluency. To learn the words would thus equal to learn the language in question. According to Nation (2001: 9), for L2 learners to reach native speaker standard they need to know very large numbers of words, close to 70,000 words, which short term and perhaps even long term might be an impossible goal. In turn, this is why each individual lexical item is not regarded as equal, at least not from a learning perspective, and why strategies of teaching and learning are so important in the EFL classroom.

Continuing on the previous idea, Nation establishes that we should be familiar with 98 % of the running words in English communicative contexts, such as books, movies or conversations, to be able to understand the content fully. Furthermore, he argues that there are around 2,000 to 3,000 high-frequency words in the English language, whereas low-frequency words could constitute several hundred thousand ones (Nation, 1990: 159). Presumably, this should also be true for languages besides English. However, this raises the question of how educators differentiate, and choose, which words are best suited for the classroom. Where exactly do we draw the line between an essential and an unnecessary word, if unnecessary words even exist? In addition, different groups of learners have different purposes for learning, making different sets of words into their learning objectives, which in turn creates a demand for different teaching methods.

In the context of this study of vocabulary frequencies, it might be worth to notice what is regarded as *knowing* a vocabulary item, or more specifically an academic word. The premise is that if the learners have used the word in writing, which is what Nation (2001: 25) describes as productive vocabulary, it is acknowledged as learnt. In relation to this matter, Schmitt and Schmitt (2005: vii) propose that knowing a word also entails frequency and register information. In other words, knowing a word includes the writer's ability to use it accordingly, with the register in mind. For the

present study, Nation's description has formed the basis for when a word is acknowledged as learnt. Any other distinction between whether the learners actually know the word, or if they use the word in the correct context, is not possible to make based on the methodology. Hypothetically, it could be easier to distinguish whether a word is known in a spoken corpus, as oral communication means separate playing rules. For instance, there is much less time to think in spoken communication, and thus errors might be easier to spot for the researcher although not for the interlocutors. Furthermore, the creator of a written production might have received feedback and later revised his or her work, which consequently means that the writing process, and the work itself, has consisted of several stages (Gilquin, 2015: 11; Myles, 2015: 313-314).

3.2 Previous Learner Corpus Research

Two crucial topics for the present study are presented in this part. First of all, the compilation, and the study, of corpora are introduced, and several examples of key learner corpora are provided. Secondly, a few relevant studies in LCR are presented, some of which are frequency-based. However, before we move on to these crucial topics, I wish to highlight this somewhat ironical quote by Geoffrey Leech: "[There] are so many corpora of potential use for English language education that it may seem perverse to suggest that they are not enough" (2011: 25). Leech proposes here that although learner corpora have risen in numbers, there are always gaps that need to be filled, and thus the selection must be expanded. Specialized corpora are in high demand, because as of right now larger general corpora have been researched extensively. Therefore, the next step must be to research corpora with narrower and narrower scope.

LCR is unique in the way that it is a discipline that incorporates theory and methods from several other disciplines, such as corpus linguistics, linguistic theory, language teaching and SLA (Prentice, 2017; Granger, 2009: 15). One of the front figures in the field of LCR was, and still is, Sylviane Granger. In 1990, Granger initiated the large international project that would become *The International Corpus of Learner English*, ICLE. The first version of it took over ten years to complete, and it is a project that is still expanding to this day. In fact, Lozano and Mendikoetxea (2013: 68) comment that "large-scale L2 corpora are rather scarce, except for ICLE", which proves its significance. The corpus contains "2.5 million words of English written by learners from 11 different mother tongue backgrounds" (Granger, Dagneaux &

Meunier, 2002: 1). In total, it consists of 6,085 essays and counting. Hasselgård and Johansson (2011: 37) suggest that it was Granger's interest in interlanguage studies, as well as her desire to expand English corpus studies beyond just NS and second language varieties, that gave birth to this significant corpus. The idea was to create an innovative corpus, which would change the course of EFL research (Granger, Dagneaux & Meunier, 2002: 1).

The ICLE has been used extensively in several LCR studies, as it is "one of the largest multi-L1 corpora (commercially) available" (Jarvis & Paquot, 2015: 614). The ICLE, however, is not without its shortcomings. Problems have been pointed out in relation to its topics and its proficiency levels (Jarvis & Paquot, 2015: 615). It may suffer from some complications, but Hasselgård and Johansson (2011: 37) also emphasize the methodology that followed the ICLE, namely *contrastive interlanguage analysis*, or CIA. This methodology incorporates two stages: native speakers, NS, in comparison with non-native speakers, NNS, as well as NNS in comparison with other NNS (Tono, et al. 2012:8). In conclusion, the key principle to take from the ICLE is that it was written by advanced English learners. For the present study, that implies that the learners were of an even higher level than the student materials that constitute the F-SCUSSE.

Consequently, many CIA studies on the ICLE have followed. Several of those studies were conducted by Ringbom (1998; 1999), who studied frequencies of words in seven Western European learner corpora incorporated in the ICLE. Ringbom, as well as Lindgren (2015), are particularly relevant for the present study, as they are some of the few who have analysed Finland-Swedish EFL learners. In fact, Ringbom, accompanied by Tuija Virtanen, functioned as the coordinator for the Finnish and Finland-Swedish sub-corpus of the ICLE. In one particular study (1998), he found that some words among advanced learners, in this case learners at university level, are particularly overused, such as personal pronouns, auxiliary verbs, vague words (*people* and *thing*), some conjuncts (*but* and *and*) and the verbs *get* and *think*, whereas words such as *the*, *this*, *these* and *by* instead are underused. At around the same time Granger and Rayson (1998) found that learners tend to overuse general and frequent nouns, such as *people*, *thing* and *problem*.

> A number of studies reveal that learners from a wide variety of (unrelated) mother tongue backgrounds display a common tendency to overuse

> common, non-specific words such as *important* (...) or *big* or *nice*.
>
> (De Cock & Granger, 2004: 78)

Ringbom was very careful about not to make any categorical statements, and instead opted to describe his findings as tendencies. According to him (1998: 51), "learners with a particular L1 *tend* to use a particular word or phrase more or less frequently than both other learner groups and native speakers", although there seems to be substantial variation even between learners from the same language background.

Ringbom's study suggests a tendency to arrive at the conclusion that L1 plays a major role in the analysis of the results. As has already been noted, and which will be discussed further in the discussion section, there seems to be a challenge to arrive at other conclusions than just the factor of L1. The role of L1 has thus been influential, although there have been other common findings. For instance, CIA research has concluded that the written English of learners is "greatly influenced by the informal spoken language" (Hasselgård & Johansson, 2011: 40). We can assume that this is because learners typically come into contact with the spoken register more often than the written register on a daily basis. This hypothesis in connection with a lack of writing experience in the target language might then mean that features from the spoken language become ever so apparent.

Similarly, Lozano and Mendikoetxea (2013) compiled the CEDEL2 (*Corpus Escrito del Español como L2*), which is a corpus supposed to represent natural learner and native Spanish, and which consists of a large sample of a variety of learners and L1 speakers. In 2015, it consisted of 750,000 words, and Gilquin (2015: 22) describes the corpus as "well-designed" and "carefully constructed". The reason for this is that the corpus has been "designed according to ten standard corpus design principles" (Lozano & Mendikoetxea, 2013: 90). These design principles relate to matters of, for instance, homogeneity, representativeness, sample size and topic. The term *homogeneity* refers in this case to a control of the texts included in the corpora, where radically different texts need to be excluded (Lozano & Mendikoetxea, 2013: 82). Furthermore, in relation to previous sections and the theme of natural language, they highlight their corpus as "a reliable source of naturalistic data" (ibid.). Once again, these principles emphasize the significance of careful corpus design in the field of LCR.

Lexical issues in relation to L2 English writing have been the focus of several studies. Usually, the focal points have been on a word class (Granger & Paquot,

2009; Schmitt & Redwood, 2011; Pípalová, 2015; Doró, 2015). The study conducted by Doró (2015) dealt with lexical measures of advanced learners in Hungary using longitudinal data, or argumentative English essays written at least six years apart. Her hypothesis was that the more recently written essays would be lexically more varied than the earlier essays. Doró (2015: 71-72) concludes that the hypothesis was confirmed, although only partially, and that the more recently written essays had a high frequency of the words *can*, *really* and *think*. In her opinion, this makes the texts appear less convincing and more speculative. Interestingly, another one of her concluding remarks is that "small local corpora built from student essays […] can serve as a basis of teaching materials" (2015: 72). In the context of the present study, this statement comes across as particularly encouraging.

In several other studies, the focus has been on sentence constructions (Virtanen, 1998; Erman, 2015; Salazar, 2014) or how learners differ from native speakers by the choice of expressions and themes (Hasselgård, 2009; Ai & Lu, 2013). For instance, Hasselgård (2009) studied argumentative English essays written by advanced Norwegian learners. The essays researched were part of ICLE. She focused on thematic structures as well as grammatical and stylistic choices, and whether or not there were traces of any L1 interference. The sample size was rather scarce, only 15,000 words, but she discovered that Norwegian advanced learners seem to "master the grammatical structures […] but not the discourse conventions of argumentative/academic writing" (2009: 138). For these reasons, she concludes that the Norwegian students' texts include a high degree of writer and reader visibility. Hasselgård's study provides the present study with beneficial insights, even though the essays were written by advanced learners and her methodology differs greatly from mine.

The fields of learner corpus research and SLA have traditionally been kept apart. Granger (2009: 14) explains that this is because universal grammar, UG, i.e. the theory of genetically coded language consisting of structural rules, has recently come to dominate SLA research. As a result, the data-driven approach has appealed little to SLA researchers. "[The] particular structure you want to investigate may not occur in natural production" (Lozano & Mendikoetxea, 2013: 67), which means that researchers in the field of SLA are forced to establish other methods. Again we encounter the recurring argument against studying naturalistic data. Another reason, as suggested by both Granger (2009: 14), and Lozano and Mendikoetxea (2013: 67), could

be that a majority of learner corpus researchers have a background in corpus linguistics or teaching, rather than in the field of SLA. This argument also goes the other way around, where most SLA researchers have been trained in quasi-experimental methods rather than in corpus methods. However, the line between the two fields becomes ever finer, as SLA researchers have started to realise the advantages of LCR and its frequency information, in their own field (Gries, 2015: 173). Similarly, LCR researchers have started to implement the methodologies of SLA into their studies. That is why *corpus-informed* studies have started to appear, i.e. studies that mix the use of corpus data and traditional SLA methodology (cf. Schmitt & Redwood, 2011). Obviously, both fields have their advantages and a combination of different types of learner data can be beneficial.

The studies may have increased in number, but nonetheless Gries (2015: 175) criticises researchers for not making their own works available enough to others. The researchers that have conducted the studies tend to not give sufficient background data or they simply do not describe the methodology in enough detail. Thus, any follow-up analyses or replications of the studies become troublesome, sometimes even impossible, to perform. This is why Gries urges researchers to "provide at least summaries of observed data" (2015: 176). A careful summary of the F-SCUSSE is supplied, to avoid any possible misinterpretations in the context of the present study. Furthermore, each step and decision are explained in detail. In terms of the methodology, the intention is to describe it in such a way that a replication of the study would be possible for anyone interested.

3.3 Academic Vocabulary

> The central role of academic vocabulary in school success is true both for native and non-native speakers of English, and at all grade levels, including primary, middle-school, secondary, and higher education.
>
> (Gardner & Davies, 2013: 305)

This quote by Dee Gardner and Mark Davies, both known for their role in learner corpus research, provides a good outset for this section on academic vocabulary. In fact, the notion that academic vocabulary is both valuable and of massive importance to learners, both native and non-native, cannot be overstated. In addition, Hinkel (2003:

72) proposes that for NNS students to succeed in academic settings, their language need to resemble the NS language. Similarly, Salazar (2014: 1) suggests that prominent non-English speaking researchers often struggle with publishing their works in a world of a predominantly English academia. A discussion on the role of English academic vocabulary in academic contexts might seem like a far stretch from the argumentative essay of upper secondary school students, but the fact remains that the two contexts share several characteristics. If there were no connection between the two, the transition in level could become too great for learners to handle.

Several decades ago Fries (1945: 4) proposed that "it is necessary to decide upon a particular type [of English] to be mastered, for there is not a single kind that is used throughout all the English speaking world". With this in mind, this is where academic vocabulary appears an appealing option for teachers to focus on, even though the teaching of it has been described as "tedious" and "somewhat boring" (Hinkel, 2003: 86). Given that many of the learners at the age of 15-18, not only in Finland but also in several other countries, have plans to continue to higher education, focusing on this component of the academic register could serve to be most valuable. In addition, much academic vocabulary is not restricted to simply academic settings but can be found in for instance newspapers and literature. At the same time, it is a type of writing unknown to many. Course designers have probably recognised the invaluable tool that the academic vocabulary is, and it is definitely a path many teachers have chosen to follow. Moreover, with the previous discussion in mind, I hope we will recognise the academic genre as something not strictly textual, but in fact conceptual too.

Nation (2001: 189-191) suggests that there are at least four major reasons for the importance of academic vocabulary for learners. This vocabulary has a high frequency in academic contexts, and simultaneously, it has been found to be generally less familiar than technical vocabulary. In addition, he proposes that academic vocabulary is in fact more approachable for teachers than for instance technical vocabulary. However, as will be demonstrated in other parts of this study, academic words do not merely include the actual words but in fact the very conceptual idea of academic and argumentative thinking. Nation does comment on this topic when he proposes that "academic vocabulary […] allows the writer to generalise talk about scientific activities" (2001: 196). Zwiers (2008: 196), on the other hand, takes on a more explicit approach and suggests that academic vocabulary has to do with the maturity of learners, or the progression from the "temporal sequencing of narrative

descriptions in younger grades to the logical structures of explaining a structure in higher grades". He also comments on the fact that academic vocabulary implies more than simply an excellent grade on the paper:

> Learning to write academically gives students more than just better scores on writing tests. The thinking that happens during the writing process helps students clarify and refine their thoughts about a complex topic.
>
> (Zwiers, 2008: 195)

There is definitely an assumption that the use of academic vocabulary goes hand-in-hand with better grades, even though Lindgren (2015: 168) found no such correlation in her data. When we compare NS learners with NNS EFL learners this natural progression of writing might be smoother if, in the latter case, focus is put on the right vocabulary, although educators must consider the variables of maturity and age. For instance, there would be no point in teaching academic vocabulary to learners of EFL who are too young to be introduced to abstract ideas. On the other hand, and relating back to what Nation (1990: 159) proposes, students that are too young for conceptual ideas might also be involved in the process of mastering their first 2,000 to 3,000 words. There are probably English courses somewhere in the world where vocabulary learning does not follow this pattern. Nevertheless, it could be considered established that academic vocabulary and higher education fit together. Furthermore, when we move beyond higher education, the thinking and communication skills associated with this vocabulary are highly regarded in many professions as well (Zwiers, 2008: 195).

In terms of corpus linguistics, and especially learner corpus linguistics where the focus is on the study and application of academic word lists, there have been many projects in the past years. Interestingly, Mark Davies and Dee Gardner, the creators of the Academic Vocabulary List of English, AVL, both function as professors at Brigham Young University, where several scholars have researched topics in relation to academic vocabulary. These include Hernandez (2017), who compared the AWL and AVL in textbook materials from an intensive reading program, and Newman (2017), who examined the occurrences of the same academic vocabulary lists in a textbook corpus. Both considered the AVL the better option because it provided better coverage. The researchers might have been slightly biased as the creators of the AVL

functioned as their professors, but their experiences of implementing the academic vocabulary lists in corpus studies have proven valuable for the present study. In addition, the implementation of the AVL in uncharacteristic settings serves as an indication of the need to investigate the coverage of the relatively new list throughout many different genres.

When Granger and Paquot (2009: 98) analysed which verbs most commonly appear in academic contexts, they found 106 verbs that "largely consist of verbs that are typically used to serve organisational or rhetorical functions prominent in academic writing: reviewing the literature, describing research, exemplifying, reporting and quoting, expressing cause and effect, describing tables and figures, contrasting and summarising". These verbs include, for instance, *associate*, *consider*, *establish* and *represent*, which subsequently should appear familiar to anyone involved in academic writing. The way in which these verbs are described by Granger and Paquot might in itself function as an outline of the academic writing as a whole.

Lindgren (2015), who also worked on Ringbom's project, studied the frequency of academic vocabulary, as well as readability in MA- and BA-theses written by Finland-Swedish students majoring in English. The corpus studied was the BATMAT corpus, where she distinguished between linguistically-oriented and literature-oriented theses. Her findings indicate that the usage of academic vocabulary in theses, as indicated by the *Academic Word List* (AWL; Coxhead, 2000) and the *New Academic Word List* (NAWL; Browne, Culligan & Phillips, 2013), differs not only due to variance in level between undergraduates and postgraduates, but also due to interdisciplinary reasons (Lindgren, 2015: 164-165). Postgraduates used a more academic language than undergraduates did, and those who wrote linguistically-oriented theses used a more academic language than those who wrote literature-oriented ones. In other words, MA writers studying linguistic phenomena appear the most prominent users of academic vocabulary. Nevertheless, this would be expected as signs of register-specific vocabulary use (Lindgren, 2015: 164). In total, the AWL accounted for a mean token coverage of 6.9 % in her materials (6.8 % for BA level writers; 7.0 % for MA level writers), whereas the NAWL only accounted for 2.2 % mean token coverage (2.1 % for BA level writers; 2.3 % for MA level writers).

In conclusion, learners should master the top two or three thousand words, as suggested by Nation, before they move on and focus on topic-specific words. The academic word lists used in this study have been constructed with a similar

pedagogical mind-set, at least the AWL, and partially the AVL too. Academic word lists, which are introduced in the next section, may be highly useful if the choice is made to focus on academic vocabulary.

3.4 Academic Word Lists

> [Academic word lists] are useful in establishing vocabulary learning goals, assessing vocabulary knowledge and growth, analyzing text difficulty and richness, creating and modifying reading materials, designing vocabulary learning tools, determining the vocabulary components of academic curricula, and fulfilling many other crucial academic needs.
>
> (Gardner & Davies, 2013: 306)

If we want to focus on academic vocabulary in a learning environment, there may be issues involved in choosing which words to include and where to find them. As previously discussed, academic vocabulary does occur in other contexts, but not close to the same extent as in academic contexts. Intermediate learners might not be too eager to read, for instance, research articles. Therefore, academic word lists, which have been around for several decades now, may function as a solution to that dilemma. As suggested by Leech (2011: 14), academic word lists are not simply useful for learners, but especially for those who prepare teaching material, including teachers and text material designers. Gardner and Davies (2013: 306) acknowledge that the compilation of academic words began in the 1970s, although it would take some fifteen years before Xue and Nation (1984) put together the representative *University Word List,* (UWL). The UWL was in fact a compilation of several preceding word lists. The list would turn out to become the new standard, and it would be used for the succeeding fifteen years. However, the mixture of words compiled from several studies suffered from inconsistency and non-representability, which meant that there was a need for an academic word list compiled of data from a large corpus of academic English (Coxhead, 2000: 214). The need for a new representative list led Averil Coxhead to begin the compilation of a list with a methodology different from previous lists. In this instance, Coxhead started to compile an academic word list, but first she excluded the top 2,000 most frequent items in the English language (2000: 213), according to the *General Service List* (GSL) (West, 1953). The large corpus of academic English, from

which Coxhead would derive the academic vocabulary, consisted of 414 academic texts. Consequently, and to the delight of many educators and researchers, the AWL was created.

> Thanks to Averil Coxhead for providing us and teachers everywhere with a principled word list to guide our teaching of academic vocabulary.
>
> (Schmitt & Schmitt, 2005: iii)

Schmitt and Schmitt are not the only ones who give praise to the list. Nation (2001: 12) describes the AWL as "very important for anyone using English for academic purposes". Coxhead states that in contrast with the UWL, "the AWL, though smaller, gives a better return on learning (2000: 226). Later, academic word lists, such as the *Academic Vocabulary List*, AVL, and the *New Academic Word List*, NAWL, would follow. We do however notice a trend of creators giving praise to the lists of their own making, such as Gardner and Davies who describe the AVL as "the most current, accurate, and comprehensive list of core academic vocabulary in existence today" (2013: 327). We also notice that the researchers who produce academic vocabulary lists share the same motivation, which is crucial for the continuing development of the lists. The intention is always to make the academic word lists better and more representative in each new version, either by the choice of a different methodology of compilation, or by the use of different source materials.

> A [word list] is only as good as the corpus it is based upon, and every corpus has limitations. Firstly, no corpus can truly mirror the experience of an individual person; rather it is hopefully representative of either the language across a range of contexts… or of a particular [domain] of language.
>
> (Schmitt, 2010: 67)

It is important to remember, just as Schmitt points out, that word lists based on a corpus or several corpora may suffer from problems similar to those related to corpora and corpus design. A representative corpus may form the basis for a well-constructed word list, although it is certainly not a guarantee. In addition, academic word lists have lately received their fair share of criticism. Certain words may be more recognised than others, but "core academic words that provide useful coverage across a range of different academic disciplines have been questioned on the grounds that such words

may change meanings when they cross those disciplines" (Gardner & Davies, 2013: 310, cf. Granger & Paquot, 2009). In other words, academic words might have ambiguous meanings depending on the discipline in which they appear. For instance, a word that has one meaning in linguistics might mean something completely different in economics. Nevertheless, the mere knowledge of such a word is sufficient knowledge in most situations.

Before we discuss the difference between the AWL and the AVL, the two academic word lists to be utilized for the present study, a couple of terms could do with some explanations. These terms include *lemmas* and *word families*. Definitions of lemmas and word families might vary, but I have opted for Paul Nation's definitions of the two terms. Thus, the distinction between lemmas and word families is best understood by an initial description of lemmas. "A lemma consists of a headword and some of its inflected and reduced (*n't*) forms" (Nation, 2001: 7). In theory, the step from a lemma to a word family is not major, but in practice, it might be troublesome as the meaning may change completely. A word family includes what a lemma includes, but in addition, it covers "its closely related derived forms" (Nation, 2001: 8). As long as the words share the same stem, they are regarded as part of the same word family. For instance, the lemma of the verb *jump* includes the forms *jumps*, *jumped* and *jumping*. Thus, all of these inflected forms constitute one lemma. On the other hand, the word family of *jump* would feature items such as *jumper* (noun), *jump* (noun). In other words, a word family can include homonyms and derived forms from other word classes.

In hindsight, many, including Gardner and Davies, have disputed the word family approach as well as the fact that Coxhead simply excluded the top 2,000 words of the English language, as proposed by the GSL (West, 1953). The GSL was already quite old at the time of the compilation of the list, and according to Gardner and Davies (2013: 309-310), as well as Paquot (2010), much academic vocabulary actually occurs quite frequently, for instance when cross-referenced with the BNC. Such frequent words that Coxhead excluded were, for instance, *business* and *exchange* (Gardner & Davies, 2013: 309). Thus, an exclusion of the top 2,000 most frequent words in the English language might not have been the most viable method according to them. Coxhead herself has even described this methodological issue concerning the GSL as a controversial decision (2011: 355). Moreover, Gardner and Davies were not convinced of the legitimacy of the word family division used by Coxhead. For instance,

the word family of the headword *react* includes words such as *reactionary* and *reactivation*, which demonstrates the fluidity in core meaning. Relating back to Nation (2001: 8), these words would typically not be described as closely related derived forms to the headword, even though they form a word family. Subsequently, these were some of the reasons why they chose to compile a new list from a corpus "both significantly larger and more recent than the corpus that was used for the AWL" (Gardner & Davies, 2013: 312). To deal with the problem of word families, Gardner and Davies divided the AVL into lemmas, which, as previously mentioned, are words with a common stem, only distinguished by inflection (2013: 308-309). The problem with the lemmas was that they did not meet the academic needs for research, which subsequently meant that Gardner and Davies were forced to form word families. However, "the key here is that lemmas, not word families, were used to make initial counts and analyses" (Gardner & Davies, 2013: 325). By the use of this approach, they have avoided this particular compilation challenge. Furthermore, both the AWL and the AVL have in common that their creators hope to have contributed to the development and improvement of EFL education. For instance, Gardner and Davies (2013: 325) express a desire to provide the "learning, teaching, and research of English academic vocabulary" with a valuable tool.

Lindgren (2015: 156) highlights that the words incorporated in academic word lists do appear in other contexts, but they are particularly frequent in the academic register. Through the years, the words have simply been acknowledged to the extent in the broad academic context that they have become appropriated, and thus standardized. In other words, the words might not be connected to a single register or field, but instead they construct a cross-academic genre (Coxhead, 2000). Argumentative writing could be a representative of texts where academic words do appear in other contexts. It should therefore be noted that argumentative writing and academic vocabulary are not presented as synonyms in the present study even though the two have formed the basis for it. However, they are definitely connected, meaning that argumentative writing could be said to constitute a sub-genre to academic writing, where writers of the former utilize similar techniques considered essential for writers of the latter.

3.5 Upper Secondary Schools in Finland

It has been my intention to look at the materials without any preconceived notions, but we must not treat the texts as disconnected from their social and linguistic context (cf. Scott and Tribble, 2006: x). Therefore, a short description of general upper secondary education in Finland is needed to put the present study, and especially its materials and the producers of those materials, in context. The focus in this part is on Finland-Swedish schools, as the aim of the F-SCUSSE is to be a representative of Finland-Swedish learner English. The differences between the English education in Finland-Swedish and Finnish upper secondary schools in Finland are likely to be minor. Only the language of teaching, besides English of course, and usually, the students' linguistic background may differ, although that is however more than sufficient to talk about the two distinct school systems in their own contexts.

3.5.1 Description

The upper secondary school in Finland is a theoretical subject-based three-year, or in some cases four-year, continuation after comprehensive school. As for Finland-Swedish upper secondary schools, in 2012 there were 37 in total and 7,077 students were enrolled in these schools. It is not compulsory to continue to upper secondary school, or to practical vocational education, but most pupils opt to do so. According to the Finnish educational board (*Utbildningsstyrelsen,* 2014: 232pp), in 2012, 36,000 Finnish students, of whom 2,300 chose a Finland-Swedish school, continued on to upper secondary school. In contrast, 50,000 Finnish students opted for vocational institutions and apprenticeship training, and 2,100 of those were Finland-Swedish students. Thus statistically, a higher proportion of Finland-Swedish students opt for upper secondary school when compared with their Finnish counterparts. A small percentage chose neither and, for instance, went on to work immediately after comprehensive school. In conclusion, a considerably high number of students continue to upper secondary school, and many of them eventually aim to get into university.

When it comes to English, most upper secondary school students choose to study English as their L2 at an advanced level, which is referred to as syllabus A. This is shown in the statistics for year 2013 by the large number of students who studied English at advanced level, 30,099 students, compared to those who studied English at an intermediate level, only 46 (*Utbildningsstyrelsen*, 2014: 235). For Finland-Swedish students the corresponding number was 2,173 students at advanced level and only 3 at

intermediate level. The intermediate level is also referred to as syllabus B. There are two major ways in which syllabus A and syllabus B in English differ from each other: mandatory courses and target levels. Syllabus A consists of six mandatory courses and two optional courses, and the target level is B2.1 as per the guidelines of CEFR. Syllabus B, on the other hand, consists of five mandatory courses and also two optional courses, and the target level is B1.1.

The central importance of English in education in Finland cannot be overstated. There have even been talks about allowing for the possibility to do every subject as part of the matriculation exam in English, except of course the language subjects (*Ministry of Education and Culture*, Finland, 2018). Previously, only Swedish and Finnish were acknowledged as acceptable languages. Presumably, this is another way of structuring the studies so that students gain even further international competence. In addition, studies have shown that Finns in general consider English the most important foreign language, even more important than Swedish, but not as important as the domestic language, in this case Finnish (Leppänen et al., 2011). Unfortunately, Leppänen et al. (2011) have not made any distinction in their results based on the variable of L1. This means that it is not possible to conclude whether Swedish-speaking Finns actually agree with this notion or not. This may only be speculation, but the assumption would be that Finland-Swedish people would have even more positive attitudes towards English, given that Swedish and English are both part of the Germanic language family. Nevertheless, in conclusion, English is a popular subject in upper secondary schools, and many students opt to study it, perhaps because the population in general find it highly useful, and because it is close to being a criterion for admission to higher education.

3.5.2 The Matriculation Examinations

In the Matriculation Examinations of English, the student may choose to write the exam either at the A-level, also referred to as *long* English, or the B-level, also referred to as *short* English. These levels are connected to the previously mentioned syllabus selections. Between the years of 2013 and 2015, 95 % of those who passed the Matriculation Examinations had chosen to write their test at the A-level. In turn, 97 % of those students accepted into university actually wrote the English Matriculation Examinations at the A-level (Pursiainen et al. 2017). This goes to show that a good

knowledge of English is valuable, and almost a given, both in the university application process as well as the studies at university.

The Matriculation Examinations are held biannually, once during the spring and once during the autumn. *Studentexamennämnden* (=the board of Matriculation Examinations) is the institution in charge of scheduling and organizing the Matriculation Examinations. Typically, a student will partake in this exam during his or her third year of study. That year the student is referred to in Swedish as *abiturient*, which roughly translates into matriculation candidate. Beginning in the spring of 2018, the English test as part of the Matriculation Examinations will become digitalised, or written on computers. The digitalisation of the English test is part of a gradual schedule that started in 2016, according to which by the spring of 2019 all subject tests for the Matriculation Examinations should have become fully digitalised (*Studentexamensnämnden*, 2018). This has led many teachers, including the ones that have supplied me with my material, to implement the application of computer writing in the classroom. There has been an increase in computer software programs, and online websites that are designed to prepare the students for the final test, such as *Abitti*. However, at this moment there are also restrictions, which means that, for instance, the use of spell-checkers and the internet are prohibited. The whole process of digitalisation is controversial and has been criticised by both students and teachers for happening too suddenly, as many of the students have gone through the educational system with pen and paper. Of course, digitalisation has its advantages for students and teachers alike. In addition, digitalisation means that materials previously uncannily tiring to compile have become more available to researchers, as long as permissions are obtained.

Previously, the English test was held on two separate days and was divided into two sections: listening comprehension, which was a 30-minute test where the student used headphones to listen to what was discussed and then answered questions, and textual production, which was a six hour test that measured reading comprehension and essay writing. Now however, technological tools have enabled the two sections to merge into one big six-hour test. The test may look different but it still measures the same skills, and essay writing remains an essential component of the overall test score, weighing around a third of the total grade. The essay should most often be argumentative and it should consist of 150 to 250 words, although teachers recommend students to write a maximum of 300 words if they aim for good grades.

Usually, the student can choose between four or five essay topics. The following is an authentic example of an essay topic originating from the test in the spring of 2017:

> **Physical exercise at school**
> Write your response to this discussion forum post in an online magazine Young Minds' web site:
> *I love PE! It's my favourite subject at school. I wish we had it every day. I'm really good at Finnish baseball. It's so exciting and always a challenge. It also develops teamwork skills. Research shows that physical activity can boost self-esteem as well as reduce your risk of stress, depression, dementia and Alzheimer's disease. If exercise were a pill, it would be one of the most cost-effective drugs ever invented.*
> – sporty Finn –
>
> (*Studentexamensnämnden*, spring 2017)

Typically, the essay topic consists of a title, a short description, and in some cases a piece of text with an argument that the student is supposed to comment on or a fictional person to respond to. The introduction of computers in the Matriculation Examinations has opened up an array of possibilities, such as the use of video and audio files, both of which had been incorporated in the spring 2018 test. According to *Studentexamensnämnden* (2018), the main purpose of the essay writing is to measure the student's ability to produce text in a range of communicative situations. It could be assumed that this involves the student's argumentative ability, as most essay topics tend to be somewhat argumentative. Occasionally however, the instructions indicate a more fictive and creative approach, such as when students are left with nothing but a title, such as in the following example.

> **Ghost story**
>
> (*Studentexamensnämnden*, autumn 2017)

This not only complicates their apprehension of the task but also the teachers' assessment: how do you measure the scores of text production in such a free form?

When students prepare for their Matriculation Examinations, they normally have mock exams and the teacher supplies them with essay titles from the previous official test, to ensure that the students are well prepared. Courses outspokenly

considered prep courses for the Matriculation Examinations might include several mock exams, and thus several essay-writing tasks. In turn, this leads us to the next section, where the materials used for compiling the F-SCUSSE are presented.

4. Materials and Methods

This section presents the materials used in this study and describes the methods used for the analysis. It starts with a presentation and description of the F-SCUSSE as well as its compilation. The F-SCUSSE is the corpus that consists of upper secondary school learners' English mock exam essays. The part on the F-SCUSSE ends with a reflection upon its representability. Later, the materials used as reference corpora are presented and assessed: the two academic word lists; the AWL and the AVL, as well as the two native speaker, NS, corpora: the LOCNESS and the BNC. Furthermore, explanations are given on how all of these materials have been produced and compiled.

4.1.1 F-SCUSSE

The materials that constitute the F-SCUSSE were collected towards the end of the autumn term of 2017. During the autumn, I contacted several teachers who work in Finnish-Swedish upper secondary schools in Finland. The four teachers who agreed to partake in this study suggested that they would contact me around the Christmas period, as this would be the point at which they themselves would receive most of their students' recently written essays. An agreement was reached, which stated that the teachers would either send the essay files per e-mail, or share them with me digitally through cloud services. Technical solutions mean that the procedure of data collection is much smoother than in earlier learner corpus compilation.

The initial procedure varied between schools, districts and teachers. Some teachers did not find it necessary that I contact the head of education in their district, whereas others recommended it. Thus, a consent form was constructed and provided to the English teachers. The consent forms were to be filled in by the parents of minors or the students themselves if they were over 18 and considered adult (see Appendix B). In one school, I handed in an application for a research permit to the head of education in that municipality (see Appendix A). In the latter instance, an agreement was reached with one individual teacher not to use any consent forms but instead she would inform the students, who in turn would agree verbally. In conclusion, all possible means were used to ensure that the students, as well as the schools and teachers, would be made aware of the research and feel comfortable with the data collection procedure. In addition, anonymity was given as a guarantee.

Beginning at Christmas 2017, and continuing through February 2018, I received more than 250 essays that contain circa 70,000 words. The four teachers, who

were willing to partake in this study and supply the materials, work in three different Finnish-Swedish upper-secondary schools in Finland. The location of the students range from the Western coast known as Ostrobothnia, to the Southern coast in the South of Finland. Thus, the sample conforms, not entirely but to a satisfactory degree to the Finland-Swedish population in general, who tend to live in those areas. To my knowledge, all of the students have Swedish as their L1, although some might be bilingual in Finnish and Swedish. Most importantly their language of education is Swedish.

There were no problems with the form of the essays, or what Granger (2002a: 8) describes as "[to] qualify as learner corpus data the language sample must consist of continuous stretches of discourse, not isolated sentences or words". Each individual text was of adequate length, and was structured with a clear start and finish. It should however be noted that the essays received were written by students at two levels in upper secondary school: first-year students and third-year students. The latter can be referred to as matriculation candidates. Ages may vary, but first-year students tend to be 16-17 years old, whereas third-year students tend to be 18-19 years old. Therefore, the materials are analysed, on the one hand, as one corpus and on the other hand as two corpora according to the level of the students.

The material was in digital form, because it had been written on school computers or private computers in the classroom by the learners. A wide range of several word processors had been used for this process, including Google Docs, Microsoft Word and Microsoft Word Online. As previously mentioned, the official word limit for a Matriculation Examinations essay in English is between 150 and 250 words, although teachers recommend students to write as much as 300 words if they aim for good grades. The mean length of these essays of circa 270 words suggests that students do aim a little above the official word limit. Now the essays totalled 67,000 words. However, the 67,000 words still include essay titles, and after these titles were excluded, the total added up to approximately 63,000 words, or tokens, consisting of 4,906 distinct words, or types. Originally, the sample size was even larger, but a number of texts had to be rejected because they were considered too fictional and not argumentative enough. In other words, homogeneity was of great importance.

It should be made clear that the students had not received any feedback on their work, nor had the essays been revised. The versions received could thus be described as representations of the writers themselves only, free from any teacher

influences. In addition, no spell checker had been used as part of the exercise for the final exam, which, as previously mentioned, was prohibited. In order to study the vocabulary of the essays in WordSmith Tools, spelling errors had to be corrected, because the software does not recognise different variants of a word as one item, unless specifically told so. However, as the F-SCUSSE corpus is relatively small, it was possible to implement a manual method of error tagging and error correcting. Common spelling errors included, for instance, problems with *th*-spellings, such as "healty" (*healthy*), troublesome words, such as "excercise" (*exercise*) and wrong use of spacing, such as in "aswell" (*as well*). By carefully reading every text, I focused on the words underlined in red by Microsoft Word, but occasionally words that were not underlined were corrected. These instances occurred when the spell-checker in Word did not acknowledge a spelling error as an error, because it had taken the form of another acceptable word, for instance *from* instead of *form.*

I made no structural changes to the texts, either grammatical or any that concerned word order. The only instances where words were changed completely were when there had been an interchangeable use of *a*/*an*. Such erroneous use of the article included, for instance, *an unit* instead of *a unit*. In addition, the preordained essay titles were marked with an opener: <, and a closer: >, so that WordSmith Tools would know to ignore them. This proofreading and error correcting was a time-consuming although necessary process, which at the same time gave me an even deeper understanding of the texts. The finalized text files were later converted into .txt files because it is one of the file types supported by WordSmith Tools. The original text files were saved in case of any need for later comparison.

In conclusion, the F-SCUSSE is a corpus that consists of argumentative learner essays and at the moment, it contains 63,000 words. It represents two groups of Finland-Swedish learners: first-year EFL learners at upper secondary school and third-year EFL learners at upper secondary school. As previously mentioned, the ages may vary, but the learners tend to be in their late teens. Furthermore, the learners originate from three different Finland-Swedish upper secondary schools in Finland. Possible spelling errors were corrected in hindsight to aid WordSmith Tools in the process of running the calculations.

4.1.2 AWL and AVL

As was previously pointed out, the introduction of the academic vocabulary lists meant that teachers around the world were offered valuable tools in the teaching of academic vocabulary to their students. For the present study, the AWL and the AVL represent academic vocabulary, and they will be used for reference, to check their coverage in both the NS and the NNS corpora, even though the focus is on the NNS corpus. Both lists were downloaded and converted into a simple .txt file where one line was equivalent to one lexical item. Similarly, as with the F-SCUSSE, this was done because it was one of the file types supported by WordSmith Tools. The AVL is available in two versions: one constructed of word families and one that consists of lemmas, or *core* academic words (Gardner & Davies, 2013a). The latter, which is a spreadsheet that contains the 3,000 most frequent core academic words as derived from the *Corpus of Contemporary American English*, COCA, is roughly the same size as the AWL list and was therefore used. Hartshorn and Hart (2016: 84) suggest that "[the] processes that underlie the development of the AVL may provide some advantages". For clarification, Gardner and Davies (2013a) describe the classification of core academic words in the following way:

> To be considered a "core academic word", it must:
> 1) Occur at least 50% more frequently in the academic portion of COCA than would otherwise be expected […]
> 2) Have a good "dispersion" […] measure of at least 0.80
> 3) Have at least 20% of the "expected" frequency in at least seven of the nine [academic] domains
> 4) Not occur more than three times as much as "expected" in any of the nine domains

The dispersion value they mention is discussed in closer detail in section 5.4. Methods. However, the downloaded AVL list also contained several duplicates of words, but these were manually deleted so that a single form of each word remained. These duplicates, or homonyms, originated from forms of words being part of many word classes, such as *value* (noun, verb), *focus* (noun, verb) or *approximate* (verb, adjective), and thus they were listed several times. The AWL was only available in one version,

which contained all items necessary, and no homonyms, and thus no adjustments were necessary.

The AWL contains 570 word families and the AVL contains at least 2,000 word families. However, when we discuss the AVL it is more purposeful to relate to the list in terms of core academic words. As such, the two lists are almost identical in terms of length, both totalling approximately 3,000 words. The AWL was constructed on the basis of a corpus that consists of 3.5 million words of academic texts published mostly in New Zealand (Coxhead, 2000: 217). The AVL, on the other hand, was compiled on the basis of a huge corpus that consists of more than 120 million words. All of the academic texts included in this corpus were published in the U.S. (Gardner & Davies, 2013: 313). According to a comparison in WordSmith WordList, the AWL and the AVL share approximately 870 words with each other, which corresponds to circa 30 % of their total type inclusion. The number is lower than one would have expected, especially when we consider that the aim of the lists is arguably the same. Averil Coxhead describes the AWL's coverage in the following words.

> The AWL contains 570 word families that account for approximately 10.0% of the total words (tokens) in academic texts but only 1.4% of the total words in a fiction collection of the same size.
>
> (Coxhead, 2000: 213)

We notice that there is a supposedly drastic difference between the presumed coverage of the genres fiction and academic texts. Gardner and Davies describe the AVL's coverage in the following way:

> [The AVL] covers ~14% of academic materials in both COCA (120 million+ words) and the British National Corpus (33 million+ words).
>
> (Gardner & Davies, 2013: 305)

For further comparison, Gardner and Davies suggest that their list covers 3.4 % of the genre fiction. Thus, the AVL has a supposedly higher coverage for academic texts, but also a higher coverage for fiction.

Gardner and Davies propose that "[the AVL] must be tested against both academic and non-academic corpora, or corpus-derived lists, to determine its validity

and reliability as a list of core academic words" (2013: 312). The F-SCUSSE is not an academic corpus per se, but its writers aim to better their academic writing skills, which places it somewhere between the genre of academic and non-academic texts. Given the status of argumentative essays, the F-SCUSSE could be said to constitute a sub-genre of academic corpora, even though the essays produced are short when compared with academic work. Nevertheless, all testing of these lists is undoubtedly advantageous for any prospective research on academic word lists.

There was a methodological issue in relation to some of the nouns included in the AVL. As will be discussed later, the methodology used has not incorporated any word class tagging, or POS-tagging (*part-of-speech tagging)*. In fact, the methodology of LCR in the area of EGAP has usually included manual error tagging (Flowerdew, 2015: 469), which could be perceived as a way for the researcher to increase the control of the texts. Therefore, some words may be misinterpreted by WordSmith tools. In this case, to call it a misinterpretation is not correct per se, because WordSmith does what it has been instructed to do. The fault simply lies in that the corpus, as well as the word lists, are yet to be tagged. In the AVL, the word *need* is only listed as academic as long as it is represented as a noun, whereas the word *need* in the F-SCUSSE would be included by WordSmith Tools as both a noun and a verb. If word class tagging is not an option, the only possible solution would be to go through the occurrences of such words manually. A concordance was manually produced for the words *need* and *use* to demonstrate this issue. *Need* occurs 109 times in the F-SCUSSE, but only 11 times as a noun, whereas *use* occurs 63 times in the F-SCUSSE, and 14 times as a noun. Thus, the concordance procedure reveals that these types of occurrences tend to lean towards verbs quite heavily. In this case, these words are not part of the academic vocabulary included in the AVL when considered verbs. This should be kept in mind when we proceed to examine the results. There are a handful of words that follow a similar pattern, for example *change*, *waste*, *value*, and *transport*. Similarly, there are words that conform to an opposite pattern, i.e. words that are listed as verbs but could be interpreted as nouns, for instance *view*. Furthermore, there are words that are listed as both verbs and nouns, for instance *value*. It is easier to avoid these challenges when we work with the AWL, as any troublesome words, such as *research*, are included as all possible word classes in the word family, for instance as a verb and a noun.

In conclusion then, it should be noted that it is the word family version of the AWL, and the core academic word list of the AVL that are used for comparison. Nonetheless, in some cases the core academic word list of the AVL will be expanded through the utilization of word families. If this is the case, it will be noted. We notice that a comparison between these two lists could prove to be problematic (cf. Hernandez, 2017; Newman, 2017), but as the lists are used to compare NS and NNS corpora, it is seemingly not a concern as long as parallels are drawn to this matter in terms of any eventual results. However, it should be highlighted that this study also sets out to measure the usefulness, or the convenience of the practical implementations of the academic vocabulary lists. Consequently, this means that such issues will be taken into consideration in the succeeding discussion.

4.1.3 LOCNESS and BNC

In this study, two distinct corpora will be used as reference corpora. The first, the *British National Corpus*, BNC, is a 100 million word corpus that includes samples of written and spoken language (approximately 90 % written data and 10 % spoken data) and which is "designed to represent a wide cross-section of British English from the later part of the 20th century" (*BNC*, 2015). It ranks as one of the largest corpora of the English language in the world. In presentation of the BNC, four specific words are used to describe it: *monolingual*, *synchronic*, *general* and *sample*. The BNC only contains British English, which categorizes it as monolingual, and it covers a period of the late twentieth century. As the BNC does not include the language's historical development, it means that the corpus is synchronic. Furthermore, the corpus is not specialized, and it involves a wide range of different genres, which categorizes it as general. Lastly, *sample* refers to the process of compilation, where the creators opted for a sample of a maximum of 45,000 words for every text. However, shorter texts can be included in their full form. The BNC will be used for reference, because it represents actual usage of English in the country mostly associated with the language. However, I oppose the notion that this corpus, as well as LOCNESS, be considered *standard*, or *correct*. Here I follow Aston (2008: 350 in Möller, 2017: 130), who questions whether native-speaker texts or "native-speaker student essays […] constitute a model to imitate". In addition, just like Möller (2017: 130), I do not wish to use a prescribed language notion as a model for what English should incorporate, but nonetheless the BNC does represent, and arguably *is* a representative of the British variety of the English language.

To compare a learner corpus with an extensive general reference corpus might not be enough. “A key facet of learner corpus research is that the learner corpus is usually compared with a native-speaker control corpus which parallels the learner corpus across as many parameters as possible” (Flowerdew, 2015: 469). In the context of the present study, the second corpus to be used as reference in this study is the *Learner Corpus of Native English Essays*, LOCNESS, which is a corpus that includes, just as the name suggests, essays written by NS learners. The learners are A-levels students of British origin, and undergraduate students of American origin. The corpus consists, at the present moment, of circa 324,000 words. However, I concentrate only on the mock exam A-levels essays, which in turn total approximately 60,000 words, or tokens, and which include 6,205, distinct words, or types. Thus, the A-levels essays in LOCNESS contain around 1,300 more distinct words than the F-SCUSSE despite the latter being 3,000 tokens smaller. In fact, it is exceptionally close to being the same size as the F-SCUSSE given the circumstances. The learners, who have written the A-levels essays, originate from a single school in the UK, and of course, this again raises the question of representability. As previously mentioned, it is important to use a NS measurement corpus that parallels the learner corpus in as many areas as possible (cf. Flowerdew, 2015: 469). To my knowledge, there is no other corpus that would contain data which would share so many features with the F-SCUSSE as the LOCNESS, and which simultaneously could be considered a NS corpus. Thus, for the reasons explained, and for lack of a better option the LOCNESS A-levels essays have been considered the materials best suited for comparison.

The reason why I choose to focus only on the A-level essays is simply that they share many characteristics with the essays in my material. The students of the A-level mock essays and the students of the Matriculation Examinations mock essays are of the same age, at approximately the same stage in their education, and although they have separate L1s, they have similar internal motivators. Many of the learners aim to continue onto higher education. Therefore, the BNC’s function as reference is to supply this study with an overall indication about the direction of the F-SCUSSE, sort of as a macro-comparison. Despite the considerably smaller size of the LOCNESS, the data it contains is much more similar in character to the F-SCUSSE. Thus, a cross-reference between these two corpora might prove to be most revealing.

4.2 Methods

Scott and Tribble (2006) describe the theoretical approaches and practical implementations of analysis of corpora when using the software WordSmith Tools. In the first five chapters, Scott thoroughly explains the various software resources which one might want to apply, whereas Tribble, in the following five chapters, puts those resources into use, for instance, by the compilation of wordlists for various corpora, and analysis of clusters. The version of WordSmith tools utilized in their book might be older than the one I use, and newer functions might have been added since then, but the manual has nevertheless been of great help to this study.

> [There] are many different ways of thinking about words and texts […] likewise there are a number of different possible display formats – the one which one needs will depend on one's research purpose, e.g. to locate words of a specific type within a larger set, or to find out patterns and qualities of the whole set of words in the list.
>
> (Scott & Tribble, 2006: 22)

Another substantial work providing methodological guidelines for this study is *The Cambridge Handbook of Learner Corpus Research* edited by Sylviane Granger, Gaëtanelle Gilquin and Fanny Meunier. This volume provides the reader with enough background, while it offers valuable advice for the whole process of a corpus study, including corpus design, corpus compilation, corpus annotating and methodology, as well as gives suggestions on how to interpret results and possible implementations of said results.

In the beginning of this study, I apply a corpus-driven, or hypothesis-finding, approach as much as possible. As the results become clearer, it is then possible to incorporate a more corpus-based, or hypothesis-driven, approach. The two terms might seem ambiguous, and there is not always a clear line between them. The initial approach, the corpus-driven approach, is to "make minimal prior assumptions about language structure […] since the corpus itself is the data" (Callies, 2015: 22), and thus patterns, and especially vocabulary structures, should emerge on their own. In a study of individual words and not a study of phraseology, John Sinclair's name might appear out of context, but nevertheless, he (2004:10) proposes that "the first stage should be an attempt to inspect the data with as little attention as possible to theory". However,

when those patterns and structures do appear it is beneficial to move from the macro-perspective to a micro-perspective and to examine those occurrences more closely. If, for instance, an academic word constantly emerges across several texts, which makes it high in frequency, it would be enough of a reason to compare it to NS corpora, or in this case, the LOCNESS and the BNC. If it then turns out that the use of the item differs across the NS and the NNS corpora, it could be an indication of linguistic variability. Furthermore, when academic words do appear, it is also a reason to investigate those that are left out, i.e. not used by EFL learners in their writing. This will become clearer in the results chapter.

Earlier, we discussed some common challenges when we deal with LCR, including the problem of avoiding monocausal conclusions, as suggested by Gries (2015: 175). That is why it has been my intention to include not just one NS corpus, but two NS corpora as a way to review any potential conclusions critically. An additional way of avoiding monofactoriality is to analyse within the corpus itself, for instance by the comparison of different age groups with each other. It cannot be denied that the factor of L1, or the factor that the NNS are language learners, would be manifested in the results. However, even in a study in the area of LCR, there is always room for slight interpretation and speculation.

The software that is used for the frequency report, the concordance and the comparison of results is WordSmith Tools, version 7. WordSmith Tools is a "software for finding patterns in the text" (Scott, 2018) that has been used and developed for 22 years, and which is still to this day undergoing development. My purpose is to use mainly two out of its three basic functions: "Concord", for finding instances of a word or a phrase in its context, as well as its collocations; and "WordList", the tool that lists words according to, for instance, frequency, or alphabetical order. Regularly, more than one of the functions will be utilized, as my assumption is that several operations offer the most thorough results. For instance, WordList forms the basis for the tables in the results, but the value of frequency per 1,000 words can only be found by the creation of a specific concordance for each individual word. In addition, the files created in WordSmith Tools, such as WordLists, will be saved as Excel files, because the latter has a simpler user interface and smoother search function, which in turn makes the calculations more straightforward.

The tables of ranked frequencies will be arranged according to what Leech (2011: 8) refers to as *ordinal frequency*, which is "a rank frequency list, in which

words are listed in order of frequency", as opposed to what he calls *raw frequency*, which is simply a statistic of how many times a certain word appears. One can think of it as a way of putting the lexical items in context, as the number in itself might not offer much insight, whereas a comparison between two distinct items may offer certain insights. The WordList function in WordSmith Tools automatically arranges the high frequency items at the head, and is followed by "an enormous tail of *hapax legomena* (words which occur once only in a corpus)" (Scott & Tribble, 2006: 11). In this context, words that occur only once are of no interest as high-frequency items and coverage are dependent on at least two occurrences. The tables of the ten most frequent words will probably contain a majority, if not an entirety, of function words, such as the determiners *a* and *the*, and the prepositions *of* and *in*, and whereas these words serve as crucial indicators of language use, I also wish to incorporate the content words. Presumably, the spread of the content words will be much greater and dependent on several variables, in contrast to the most common function words that can take up as much as 25 % of the running words in the English language (Leech, 2011: 9). To do so WordSmith Tools must be set up to ignore certain words, a feature in the software known as *Stop List*. Stop lists are simply "lists of words which you don't want to include in analysis […] like *the*, *of*, *was*, *is*, *it*" (Scott, 2018a: 120). The list used for this analysis, which was titled "English stoplist", was included in WordSmith Tools, and has been constructed from the BNC. The list was constructed by taking the top 200 most frequent words, "excluding # [numbers] and filtering out any open class nouns, verbs, adjectives" (Scott, 2018), which in turn resulted in a list that contains 142 items.

It should be noted that I have treated each word form as an individual item, which means that in the results section, several word family members or lemmas could appear in the same list. There are some exceptional cases where the whole word family is intended, but those are clarified in connection to the tables or figures. Furthermore, the lexical items will be listed according to various measurements. The meaning of some are clear, such as *Frequency* and *Frequency per 1,000 words*. On the other hand, some might not be as clear, such as *N*, which implies the ordinal frequency, and not the as per usual the number of occurrences, and *%*, which indicates the percentage total of that corpus or the specified sub-corpus. In case of any irregularities, further explanations will be given in the table headers, or in connection to the tables or figures. In contrast, the *dispersion value* might be a term that could do with an explanation. The dispersion value is based on a mathematical formula (Oakes, 1998, in

Scott, 2018a: 446) and it ranges between 0.01 and 1.00. This value expresses "the degree to which a set of values are uniformly spread" (Scott, 2018a: 446). In the cases of the F-SCUSSE and the LOCNESS, a high dispersion value, > 0.90, would indicate that a lexical item appears evenly across several of the essays, whereas a low number would indicate burstiness, or that only a few students use it regularly. Most often, the number will be slightly lower or slightly higher than the mean value, 0.50. In some instances, for example when the two reference corpora are used for comparison with the F-SCUSSE, the dispersion value will be given for results in the BNC. However, the dispersion value in the BNC might not be as indicative, and thus comparable, as in the learner corpora, due to its very large sample size.

All of the calculations, except for the coverage statistics, have been performed by WordSmith Tools. The token and type coverage for the present study were calculated using a method developed on the basis of WordLists produced by WordSmith Tools. The first step was to make a WordList, which consisted of two files: one text file of the words of the essays and one text file of academic vocabulary from either the AVL or the AWL. In WordList, there is an option to analyse in how many texts the item appears. In this case, a "2" meant that one academic vocabulary item appeared in both files, which in turn meant that the essay texts included that item somewhere in them. The list was then exported to Microsoft Excel, and a new file was constructed. Thus, the new file with its list contained only the items marked with a "2". Now all items in this newly created list were items that were a), academic vocabulary, and b), academic vocabulary items that the students had used. It was then possible to count the sum of frequencies. However, the total number of academic words had to be subtracted from the total number of occurrences of the items, as one occurrence of each word would always be in the academic vocabulary list. For instance, the word *significantly* was marked as having occurred 3 times. This meant that the writers in the F-SCUSSE had made use of it twice, but one occurrence was in the AWL, so one occurrence had to be excluded. That is what is referred to as *false matches* in the formula. Thus, all the information needed to calculate the coverage was now available: the token frequency of words in the corpus and the frequency of all of the items in the list, the type frequency in the corpus and the sum of matches between the list and the corpus. For instance, the token coverage of the AWL in the F-SCUSSE was calculated using the following formula:

((2636 - 555) / 63,263) X 100 ~ 3.3 %
"occurrences" "(false) matches" "total tokens" "percentage"

The type coverage of the AWL in the F-SCUSSE, on the other hand, was calculated as follows:

555 / 4,885 X 100 ~ 11.7 %
"matches" "total types" "percentage"

In order to establish the viability of this methodological approach, as well as the implementation of the academic vocabulary lists, a random selection of academic articles from the linguistic discipline (see "Academic Writing Sample" in references) was collected to test the assumed word list coverage. In a sample size of a mere 10,000 words, the AWL had a token coverage of 10.6 %, and the AVL had a coverage of 14.7 %. This is exceptionally close to the proposed coverage of 10 % for the AWL (Coxhead, 2000: 213) and 14 % for the AVL (Gardner & Davies, 2013: 305), which strongly suggests that the methodology is not arbitrary and that the form of the academic word lists works well. The test sample size was rather small, but it could be assumed that a larger sample size for testing would have made the coverage numbers converge towards the proposed statistics. Alternatively, such a minor difference could relate to either the authors' lexical choices or to simply a slightly higher coverage.

5. Results

This section is divided into two parts. The first part, frequency results, covers the most frequent words in the BNC, the F-SCUSSE and the LOCNESS. In this part, the tables for the BNC, the F-SCUSSE and the LOCNESS have been separated into two components, for instance 1.1 and 1.2. The .1 components show the regular most frequent words, as indicated by WordSmith WordList, whereas the .2 components display the most frequent words when the stop list has been applied. As previously mentioned, the function of the stop list is to filter out the most common function words. The ordinal frequency in the tables ranges between 15 and 20 items. These results might not offer us much as such, but they may be an indication of whether the language broadly conforms to the NS reference corpora or not. In addition, these results function as a prelude to the second part, by serving as an indication of what constitutes NS and NNS argumentative writing when we exclude function words. The second part then presents the results of the academic vocabulary analysis. There, the focus is put on the academic vocabulary lists, and their coverage of the F-SCUSSE and the LOCNESS. The items included in the AWL and the AVL that appear most frequently in the student corpora are then presented and compared. This comparison is performed at several levels, in other words which means that corpora and sub-corpora are compared. In addition, a selection of words from either academic word list not used in writing in the F-SCUSSE are discussed. In terms of how this procedure was performed is described in more detail further ahead.

5.1 General Frequencies

In the following tables, the fifteen most frequent words in the NNS corpus and the reference corpora are ranked according to their ordinal frequency. Initially, the data displayed in the tables show the word's ordinal frequency, its frequency in that corpus, and the percentage to which that frequency translates into in the corpus. To begin with, table 1.1 illustrates the fifteen most frequent lexical items in the BNC, and consequently, perhaps even in the British variety of the English language. It is no surprise that all of these words happen to be function words, such as determiners and the verb *be* and its conjugations. In fact, function words will generally be among the most frequent items whenever we deal with frequency results in corpora, as it is quite difficult to put together a sentence without the use of them. When we look at table 1.1

we encounter words that can be concluded to be common in both spoken and written language.

Table 1.1. 15 most frequent words in the BNC

N	Word	Frequency	%
1	THE	6,055,105	6.09
2	OF	3,049,564	3.07
3	AND	2,624,341	2.64
4	TO	2,599,505	2.61
5	A	2,181,592	2.19
6	IN	1,946,021	1.96
7	THAT	1,052,259	1.06
8	IS	974,293	0.98
9	IT	922,687	0.93
10	FOR	880,848	0.89
11	WAS	863,917	0.87
12	I	732,523	0.74
13	ON	731,319	0.74
14	WITH	659,997	0.66
15	AS	655,259	0.66

The appears to be the most frequent, with a remarkable frequency of 6 % in the BNC, although several other items display prominence. In contrast, table 1.2 illustrates the words that remain when the previously mentioned stop list has been applied in the calculations.

Table 1.2. 15 most frequent words in the BNC (stop list applied)

N	Word	Frequency	%
1	TIME	152,571	0.15
2	LIKE	146,825	0.15
3	NOW	136,584	0.14
4	FIRST	124,256	0.12
5	NEW	123,229	0.12
6	PEOPLE	120,623	0.12
7	KNOW	119,076	0.12
8	SEE	114,552	0.11
9	WAY	99,778	0.10
10	MADE	91,242	0.09
11*	WORK	89,678	0.09
12	RIGHT	89,547	0.09
13	YEARS	88,611	0.09
14	THINK	88,379	0.09
15	GOOD	78,338	0.08

*= the spoken interjection, or hesitation marker, *er* was ranked 11th but not included in this ranking

The most common function words disappear, and what remains are simply common words, such as *people*, *know* and *think*. However, it could be argued that some of these words fulfil functions similar to that of function words, for instance *like* as a preposition with function as a filler, and which is homonymous with the verb. Here too we notice that many of the lexical items in table 1.2 are words that are associated with both spoken and written language. Table 2.1, on the other hand, illustrates the ordinal frequency in the F-SCUSSE. A quick inspection of the words suggests that nearly all lexical items are the same as in table 1.1, even though they are slightly rearranged. The only words that are among the fifteen most frequent in the BNC, but not in the F-SCUSSE, are the words *for*, *on*, *with* and *as*.

Table 2.1. 15 most Frequent Words in the F-SCUSSE

N	Word	Frequency	%
1	THE	2,589	4.09
2	TO	2,232	3.53
3	AND	1,808	2.86
4	A	1,547	2.45
5	IS	1,348	2.13
6	OF	1,144	1.81
7	THAT	1,126	1.78
8	IT	1,077	1.70
9	YOU	1,023	1.62
10	IN	1,022	1.62
11	I	957	1.51
12	FOR	650	1.03
13	ARE	643	1.02
14	HAVE	640	1.01
15	BE	607	0.96

When tables 1.1 and 2.1 are compared, we instead encounter the personal pronoun *you*, and the verb forms *are*, *have* and *be* in the ordinal frequency ranking of the F-SCUSSE. However, when the same stop list is applied to ignore certain function words in the F-SCUSSE, the ordinal frequency changes completely. The table is now expanded to include 20 items to demonstrate even further the diversity of the words that appear. Furthermore, the dispersion value is included for the same reasons. As in table 1.2, some function words remain un-ignored in table 2.2, such as *like*, but what remains are mostly content nouns, verbs and adjectives, although some of them may seem somewhat vague (cf. Ringbom, 1998). Already at this stage it is possible to establish

how many of these words and the words in the other .2-tables, are considered academic by manually cross-referencing with the academic word lists. These items that are part of the academic vocabulary are marked by an asterisk.

Table 2.2. 20 most Frequent Words in the F-SCUSSE (stop list applied)

N	Word	Frequency	%	Dispersion
1	PEOPLE	426	0.67	0.88
2	TIME	277	0.44	0.72
3	THINK	276	0.44	0.94
4	LIKE	244	0.39	0.87
5	GOOD	182	0.29	0.87
6	CHILDREN	180	0.28	0.37
7	WORLD	162	0.26	0.74
8	THINGS	159	0.25	0.79
9	LIFE	157	0.25	0.85
10	IMPORTANT*	149	0.24	0.75
11	LOT	146	0.23	0.86
12	WAY	145	0.23	0.94
13	DIFFERENT	140	0.22	0.80
14	NEW	138	0.22	0.85
15	MAKE	129	0.20	0.85
16	ART	126	0.20	0.51
17	SOCIAL*	124	0.20	0.65
18	TECHNOLOGY*	124	0.20	0.61
19	EXAMPLE*	123	0.19	0.93
20	WANT	119	0.19	0.81

*=academic word according to the AWL/AVL

Out of the 20 most frequent items in the F-SCUSSE, four items that are considered academic according to either the AWL or the AVL appear particularly frequently when a stop list is applied: *important*, *social*, *technology*, and *example*. However, *important* and *example* attract attention as they are indicated to have a high dispersion value, meaning they are frequently used across several text samples.

The third corpus to be used for comparison is the LOCNESS, which consists of the A-levels essays. The fifteen most frequent words in the LOCNESS are shown in table 3.1. The ordinal frequency of table 3.1 may not resemble that of table 1.1 (BNC) and 2.1 (F-SCUSSE), but if we focus on what words are displayed, we notice that many lexical items are in fact shared. In fact, just in terms of the most common items, the LOCNESS shares more words with the F-SCUSSE than with the BNC.

Table 3.1. 15 most frequent words in the LOCNESS (A-levels essays)

N	Word	Frequency	%
1	THE	3,993	6.63
2	TO	1,990	3.30
3	OF	1,854	3.08
4	AND	1,453	2.41
5	A	1,336	2.22
6	IS	1,296	2.15
7	IN	1,053	1.75
8	BE	825	1.37
9	IT	822	1.36
10	THAT	744	1.24
11	FOR	620	1.03
12	THIS	616	1.02
13	ARE	566	0.94
14	AS	564	0.94
15	HAVE	497	0.83

When we choose to ignore function words, the pattern of the previous tables (tables 1.2 and 2.2) appears to a similar extent in table 3.2, which displays the most frequent items in the LOCNESS when a stop list is applied.

Table 3.2. 20 most frequent words in the LOCNESS (A-levels essays, stop list applied)

N	Word	Frequency	%	Dispersion
1	PEOPLE	393	0.65	0.89
2	BEEF	195	0.32	0.30
3	LOTTERY	191	0.32	0.27
4	BOXING	181	0.30	0.29
5	MONEY	137	0.23	0.65
6	HUMAN	133	0.22	0.62
7	SPORT	132	0.22	0.36
8	BRAIN	117	0.19	0.66
9	COMPUTER*	112	0.19	0.48
10	GENETIC	99	0.16	0.20
11	USE*	96	0.16	0.72
12	CHILD	95	0.16	0.34
13	DISEASE	89	0.15	0.45
14	TRANSPORT*	86	0.14	0.12
15	USED	86	0.14	0.79
16	WORK	84	0.14	0.65
17	WORLD	84	0.14	0.79
18	GOVERNMENT	83	0.14	0.75
19	PUBLIC	83	0.14	0.62
20	SCIENTISTS	82	0.14	0.36

*=academic word according to the AWL/AVL

Table 3.2 is expanded to include 20 lexical items and the top occurrences include words such as *people*, *money*, *computer* and *disease*. In contrast to the F-SCUSSE, when the function words are ignored the LOCNESS contains mostly countable nouns, just a single adjective, and no verbs. Out of these 20 items, three items are considered academic according to either the AWL or the AVL: *computer*, *use* and *transport*. In fact, the F-SCUSSE shares three words with the BNC, whereas the LOCNESS only shares 1 item with the BNC when stop lists have been used in the corpus searches. Furthermore, when we examine table 3.2 it is possible to estimate which topics the students have dealt with in their argumentative essays. In contrast, it is rather difficult to determine any possible topics that the writers of the essays in the F-SCUSSE have explored when we investigate the equivalent parts in table 2.2. Thus, judging from table 3.2, the essay topics seemingly range, for instance, from the meat industry (*beef*) to one that concerns martial arts (*boxing* and *sport*). In reality the writers dealt with the topics of transport, boxing, the parliamentary system and fox hunting (LOCNESS description). In addition, the frequent use of technical vocabulary fits the genre, although the dispersion values suggest that the words are quite text-specific. However, it is remarkable that the supposedly vague word *people* appears nearly twice as frequently as the second most frequent item in both NS and NNS learner writing. The item was not expected to show up to this extent in NS student writing.

Because the word *people* appears so very frequently, the word was chosen for closer analysis. The vocabulary item was therefore checked for collocations, or frequencies of co-occurrences (Nation, 2001: 328), with the help of Concord in WordSmith Tools. The concordance search was performed both in the F-SCUSSE and in the LOCNESS, and the focus was on the word that preceded *people*. The results from these concordance searches are available in table 4. It seems that in the F-SCUSSE the word is most commonly preceded by the word *some*.

Table 4. Preceding words to *people* and their frequencies

Word	Frequency in the F-SCUSSE	Frequency in the LOCNESS (A-levels essays)
SOME	39	10
MANY	33	73
OTHER	25	-
THE	21	23
MORE	5	14

It is preceded by *some* 39 times, *many* 33 times and *other* 25 times. In the LOCNESS *people* occurs after *many* 73 times, after *the* 23 times, after *more* 14 times and after *some* just 10 times. Interestingly, there were no occurrences of the collocation *other people* in the LOCNESS A-levels essays. Figure 1 demonstrates a screenshot from WordSmith Concord of the use of the collocation *some people* in 12 of the 39 total occurrences in the F-SCUSSE. In these instances, we can see that the collocation *some people* does result in generalisations as well as vagueness, which raises questions. However, those specifics remain unanswered for the most part.

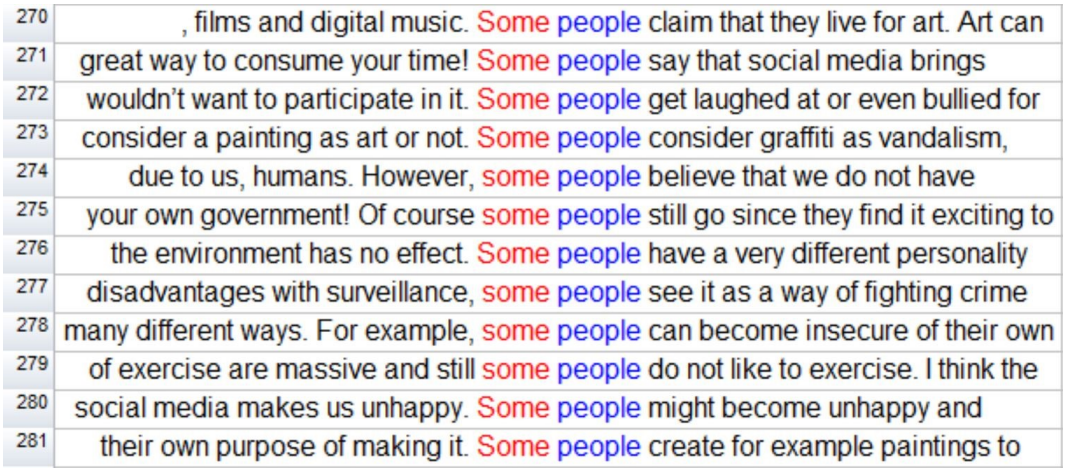

Figure 1. *People* co-occurring with *some* in the F-SCUSSE

Table 2.2 and table 3.2, which present the most frequent words in the F-SCUSSE and the LOCNESS when a stop list is applied, form the basis for table 5. The 15 words chosen for closer inspection are among the most frequent lexical items either in one of the student corpora, or in both. Countable nouns, such as *beef*, *lottery*, *boxing* and *computer* were excluded, as they were considered excessively topic-tied words, which their dispersion value confirms. In comparison, the words that have been included in table 5 could be considered more general and thus less context-bound, for instance *children*, *time*, *human* and *things*. Most general words seem to be preferred by the writers in the F-SCUSSE, although the exact reasons for this remain unknown. In contrast, the high-frequency words in LOCNESS include many topic-tied words, as previously suggested in the comments on table 3.2. As a result, a majority of the words chosen for analysis are among the most frequent in the F-SCUSSE. Thus, table 5 has been arranged according to the words' frequencies in the F-SCUSSE, and two results are given per cell: frequency per 1,000 words and the dispersion value in brackets. The

right-most column features results from the BNC for comparison. In this instance, we begin to discover an indication of word frequency variation in the F-SCUSSE when compared to the NS LOCNESS and the NS BNC. As previously mentioned, the word *people* remains an item used frequently in both the NS and the NNS learner corpora. EFL learners however seem to use words such as *think* and *like* approximately four times more often than NS learners and *good* twice as often. Furthermore, words such as *world* and *life*, and especially *things*, *important* and *lot* appear to be particularly high-frequency items.

Table 5. A selection of the most frequent content words in the F-SCUSSE and in the LOCNESS in comparison with the BNC (stop list applied)

Word	Frequency in the F-SCUSSE per 1000 words (dispersion)	Frequency in the LOCNESS per 1000 words (dispersion)	Frequency in the BNC per 1000 words (dispersion)
PEOPLE	6.81 (0.88)	6.56 (0.89)	1.28 (0.96)
TIME	4.41 (0.72)	1.09 (0.90)	1.56 (0.98)
THINK	4.41 (0.94)	1.17 (0.71)	1.00 (0.79)
LIKE	3.89 (0.87)	0.83 (0.83)	1.53 (0.89)
GOOD	2.90 (0.78)	1.21 (0.72)	0.84 (0.91)
CHILDREN	2.87 (0.37)	1.11 (0.66)	0.57 (0.93)
WORLD	2.60 (0.74)	1.40 (0.79)	0.64 (0.92)
LIFE	2.55 (0.85)	1.22 (0.75)	0.59 (0.92)
THINGS	2.53 (0.79)	0.33 (0.80)	0.46 (0.93)
IMPORTANT*	2.38 (0.75)	0.38 (0.86)	0.42 (0.92)
LOT	2.32 (0.86)	0.53 (0.85)	0.36 (0.97)
WAY	2.32 (0.92)	1.23 (0.78)	1.02 (0.97)
DIFFERENT	2.25 (0.79)	0.32 (0.70)	0.50 (0.99)
CHILD	1.86 (0.31)	1.78 (0.34)	0.34 (0.88)
HUMAN	0.53 (0.81)	2.22 (0.62)	0.26 (0.89)

*= academic word according to the AVL

The plural vague noun *things* does in fact appear almost eight times as frequently in the F-SCUSSE as in the LOCNESS. Interestingly, the only word out of these that is not among the most frequent in the F-SCUSSE is the word *human,* which could be used as a synonym, at least in some cases, for the word *people*. Other vague words with high frequencies in the F-SCUSSE include *good*, *lot* and *different*. It could be assumed that many of the words mentioned in this section would be classified as lexical teddy bears. Furthermore, as can be observed in table 5, the dispersion values for the words included are remarkably high in the BNC. Every word in the table, except the word *think*, has a

dispersion value higher than 0.88, which suggests that these words are very common across several genres, and that they are used commonly in both written and spoken language. The item *think* has a dispersion value of only 0.79 in the BNC, which is a bit surprising. Nonetheless, it would still be categorised as quite widely used. In contrast, the words *different*, *lot* and *way* are exceptionally common in the BNC with dispersion values greater than 0.97.

In conclusion the analysis of the general frequency provided more thought-provoking data than expected. There are undoubtedly differences in the lexical use of NNS learners and NS learners, but similarly there are resemblances between the two when contrasted to the BNC. The latter can be taken as an indication that the genre, that is argumentative writing, is shared. In addition, we can observe items associated with general and especially spoken language to a higher degree among the most frequent items in the essays that constitute the F-SCUSSE than in the LOCNESS A-levels essays. Occasionally however the same items do appear relatively frequently in the writings of NS learners too, which conflicts with the preconceptions of expert writing in comparison with non-expert writing.

5.2 Frequencies of Academic Vocabulary

The focus now shifts from general frequencies to academic vocabulary and the academic vocabulary lists. Most of these results were calculated using the tools and formulas described in the methods section. However, a few additional findings emerged as by-products of the data-driven approach, just as in the previous section and the case of *people* and its collocations. These additional findings are presented and explained at relevant points towards the end of the section.

This section begins with the type and token coverage of the academic vocabulary lists in each learner corpus. This is consequently why two numbers are presented in the figures, namely the type coverage and the token coverage. This leads us to figure 2, which shows that the Academic Word List, AWL, covers 11.3 % of types in the F-SCUSSE, whereas it only covers 3.3 % of the total size of 63,204 tokens. The Academic Vocabulary List, AVL, on the other hand, has a type coverage of 11.7 % and a token coverage of 5.7 % in the same corpus. Thus, the academic vocabulary lists have a similar type coverage in the F-SCUSSE but a notable difference in terms of token coverage.

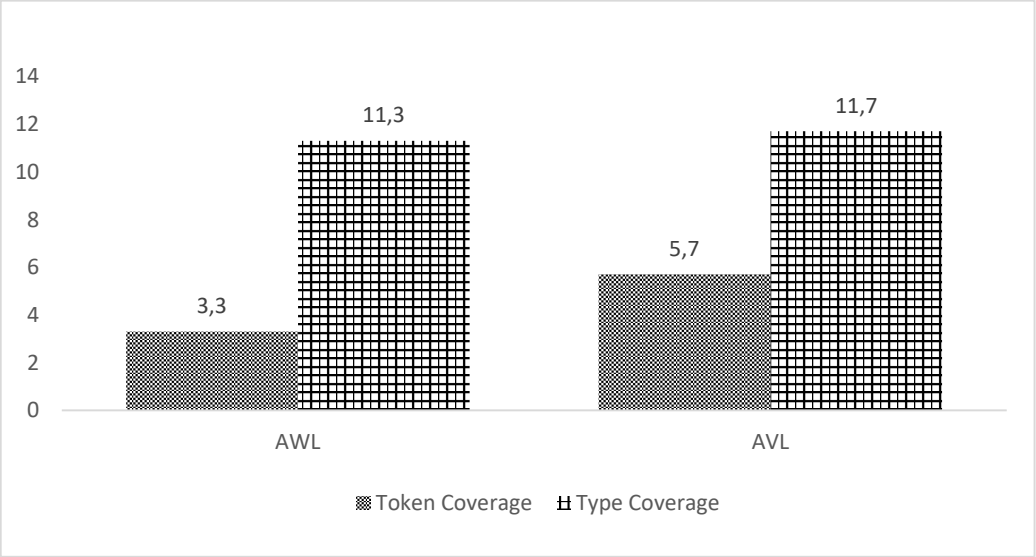

Figure 2. Coverage of the AWL and the AVL in the F-SCUSSE

For comparison, figure 3 presents the coverage of the AWL and the AVL in the LOCNESS corpus, displayed in the same manner as in figure 2. In the A-levels essays part of LOCNESS, the AWL has a token coverage of 4.9 %, and a type coverage of 12.7 %. In contrast, the AVL contributes for a token coverage of 7.5 % and a type coverage of 13.4 %. Hence, all numbers appear superior to the equivalent data in the

F-SCUSSE. As previously mentioned, the F-SCUSSE contains only slightly more data (5,000 words) than the LOCNESS, which allows for a direct comparison to be made between the two corpora. At this point, it is also worth noticing the apparent drop in coverage for the AWL in both the LOCNESS and the F-SCUSSE when the type coverage and the token coverage are compared.

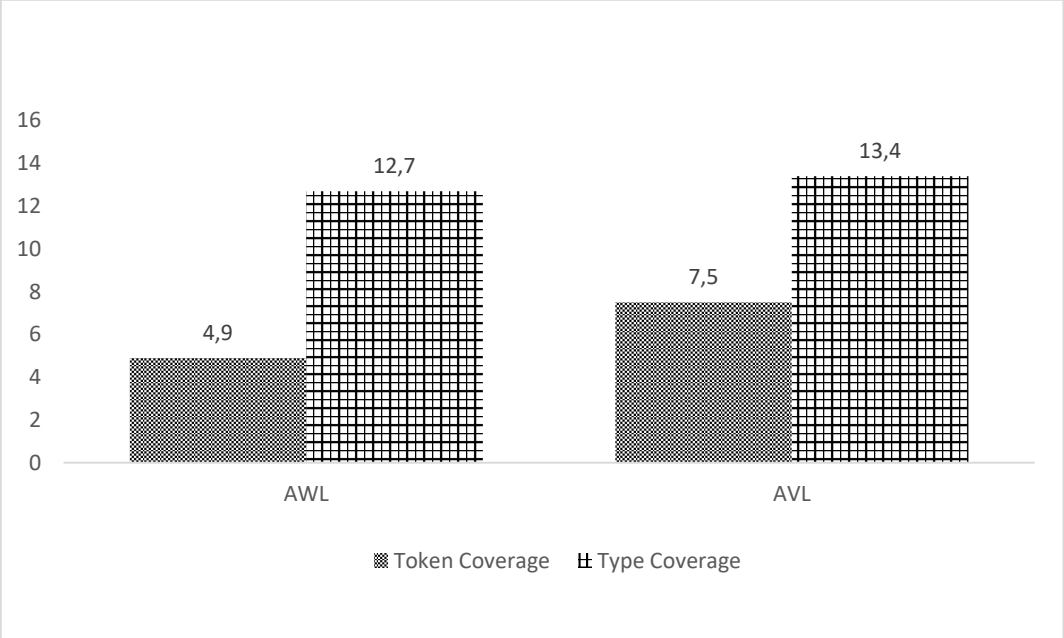

Figure 3. Coverage of the AWL and the AVL in the LOCNESS (A-Levels essays)

When the F-SCUSSE is divided into sub-corpora there is a possibility to measure the coverage of the AWL and the AVL in them. The sub-corpora are then labelled according to whether the students are first year students or third year students. However, it should be noted that the sub-corpora are of substantially different sizes. The sample size for the first year students accounts only for approximately a sixth of the F-SCUSSE, which means that the sample size for the third year students is proportionately much larger (11,000 tokens in contrast to 52,000 tokens). For figure 4, the procedure was the same as in the calculations for the previous two figures (figure 2 and figure 3). This means that it was constructed with the use of the same formula for calculation introduced in the methods chapter.

In figure 4, we can see a difference in the coverage of the academic word lists between the two F-SCUSSE sub-corpora. More specifically, it seems that third year students would use a more academic vocabulary than first year students do, if we consider higher coverage statistics to equal a more academic writing style.

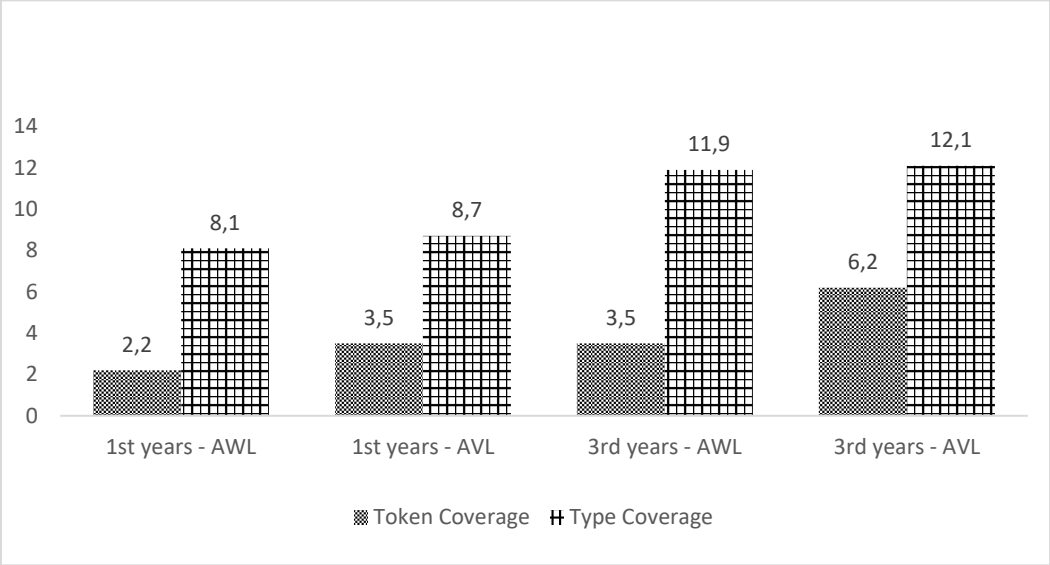

Figure 4. Coverage of the AWL and the AVL in the F-SCUSSE according to level

Interestingly, the difference between the lists' type coverage for first year students and for third year students is roughly the same irrespective of the list. Instead, the main distinction lies in the token coverage. In addition, it seems that the numbers for the third year students are positioned approximately in between first year students and NS students. In fact, it would have been arguably rather controversial if the results had pointed to the opposite. As previously mentioned however, the difference in sample size needs to be taken into consideration when analyzing these findings. Despite these reservations, the findings offer an opportunity to suggest that these results could possibly represent the pseudo-longitudinal development of Finland-Swedish learners, even though the difference in level is only two study years. However, two study years could certainly be considered a long interval in the context of classroom teaching.

The coverage statistics makes for interesting observations and there is reason to believe that similar patterns would emerge when individual words are checked. Therefore, let us now turn to the specific academic vocabulary items to be able to establish how they occur in the NS and the NNS learner corpora. All of the following tables were compiled by conversion of the results directly from the WordList files turned Excel files. Three numbers are given per word in the tables: the frequency in the corpus, the frequency per 1,000 words and the dispersion value. The ordinal frequency which decides the ranking of the words could be seen as a fourth implicit number. Thus, table 6 displays the most common academic words from the AWL in

the F-SCUSSE. These words tend to be either part of the technical vocabulary or belong to the word class of nouns, or both. Nouns that are simultaneously part of the technical vocabulary remind us of the underlying ideas that influenced the construction of the AWL. Besides nouns and adjectives, there are not many occurrences of words outside these word classes, except for the items *relax*, *found* and *create*.

Table 6. Most common academic words (AWL) in the F-SCUSSE

Word	Frequency	Per 1,000	Dispersion
TECHNOLOGY	124	2.10	0.61
MEDIA	99	1.61	0.58
NUCLEAR	56	1.27	0.52
PHYSICAL	40	0.64	0.54
AUTHORITIES	34	1.09	0.50
ENVIRONMENT	29	0.60	0.62
ENERGY	26	0.42	0.67
INSTANCE	24	0.39	0.82
RESEARCH	23	0.37	0.63
TOPIC	22	0.36	0.79
RELAX	22	0.36	0.77
JOBS	22	0.36	0.57
POSITIVE	21	0.35	0.67
COMPUTERS	21	0.35	0.48
ROLE	19	0.32	0.34
MENTAL	18	0.30	0.58
STRESS	16	0.26	0.73
FOUND	16	0.26	0.77
CREATE	16	0.26	0.62
NEGATIVE	15	0.25	0.66

Table 7 shows the most common academic lexical items in the F-SCUSSE, as suggested by the AVL. As previously mentioned, these are not word family members as with the AWL, but instead they constitute what Gardner and Davies (2013) refer to as core academic words. The key difference when contrasted with table 6 is that high frequency words of several word classes start to occur. There are linking words, such as *however*, *example* (as in *for example*) and *therefore*, i.e. words that are indeed very prominent in academic writing. The nouns *need*, *use*, *change* and *experience*, as has been discussed previously, are slightly troublesome in this context, as it is not possible to know which word class they belong to in the corpus without checking each occurrence for its word class.

Table 7. Most common academic words (AVL) in the F-SCUSSE

Word	Frequency	Per 1,000	Dispersion
HOWEVER	152	2.43	0.80
IMPORTANT	149	2.38	0.75
TECHNOLOGY	124	1.98	0.61
SOCIAL	124	1.98	0.65
EXAMPLE	123	1.96	0.93
NEED*	109	1.74	0.89
CHANGE*	81	1.37	0.68
FUTURE	74	1.20	0.62
QUALITY	72	1.15	0.37
CLIMATE	64	1.02	0.47
USE*	63	1.00	0.80
SOCIETY	48	0.78	0.89
COMMON	40	0.64	0.80
BOTH	40	0.64	0.82
MEAN*	33	0.54	0.89
EXPERIENCE	31	0.50	0.69
SCIENCE	31	0.50	0.54
ENVIRONMENT	29	0.47	0.62
THEREFORE	28	0.45	0.73
MODERN	28	0.45	0.81

*= words classified as academic only when nouns

Some of these cases were counted manually to test word class distribution. This was performed to fully understand the margin of error we were dealing with. Thus, it seems that when the marked words in table 6 are checked manually, it is apparent that the homonyms appear more often as verbs than as nouns in the learners' writing. For instance, the word *need* occurs 109 times, but merely 11 times as a noun, whereas the word *use* occurs 63 times, but only 14 times as a noun. Considering the words in table 7, this does lead to inflation in numbers for several items. It is therefore obvious that these circumstances must be taken into consideration when analysing the results. On the upside, the fact that the students have used a word family member that also happens to be a homonym to the intended word could indicate that they have knowledge of the other item which belongs to another word class.

As we saw in table 2.2 and table 3.2 in the previous section, that the two learner corpora differ in their most frequent general vocabulary, it could be assumed that they also differ in their most frequent academic lexical items. Table 8 shows the results of the most commonly occurring academic vocabulary items when the LOCNESS is cross-referenced with the AWL. A comparison with table 6 shows that

the F-SCUSSE and the LOCNESS share six items in their lists of the 20 most common AWL items: *technology*, *research*, *job/jobs*, *computer/computers*, *found*, and *create*.

Table 8. Most common academic words (AWL) in the LOCNESS (A-levels essays)

Word	Frequency	Per 1,000	Dispersion
COMPUTER	112	1.87	0.48
TRANSPORT	86	1.47	0.12
COMPUTERS	81	1.35	0.44
MAJOR	48	0.80	0.65
MANIPULATION	44	0.76	0.17
TECHNOLOGY	40	0.67	0.68
SEX	35	0.57	0.19
AREAS	30	0.50	0.76
ISSUE	27	0.48	0.81
RESEARCH	24	0.43	0.42
FOUND	22	0.37	0.64
CREATE	22	0.37	0.72
MAJORITY	20	0.35	0.72
INJURIES	20	0.35	0.26
PROCESS	19	0.35	0.42
MEDICAL	19	0.35	0.70
INDIVIDUAL	18	0.33	0.71
INCOME	18	0.33	0.64
DESPITE	18	0.33	0.65
JOB	17	0.31	0.64

Table 9 presents the results of a cross-reference between the AVL and the A-levels essays of the LOCNESS. The item *however*, which indicates an oppositional stance from the writer, stands out as the most frequent item. Out of the 20 AVL items, nine words are shared with NNS learners' vocabulary use as listed in table 7: *however*, *technology*, *example*, *need*, *future*, *use*, *society*, *both*, and *therefore*. In other words, quite a substantial number of words are shared between the two tables, which subsequently means that there are similarities between the two learner corpora. The observation that there are various similarities between the two learner corpora could be described as unexpected but simultaneously interesting.

Table 9. Most common academic words (AVL) in the LOCNESS (A-levels essays)

Word	Frequency	Per 1,000	Dispersion
HOWEVER	146	2.43	0.84
HUMAN	133	2.22	0.62
USE*	96	1.60	0.72
TRANSPORT*	86	1.47	0.12
EXAMPLE	68	1.17	0.82
THEREFORE	61	1.02	0.71
SYSTEM	56	0.98	0.43
INCREASE	51	0.85	0.70
NEED*	47	0.81	0.84
MANIPULATION	44	0.76	0.17
EFFECT	42	0.70	0.84
TECHNOLOGY	40	0.67	0.68
INCREASING	37	0.62	0.45
SOCIETY	35	0.64	0.76
GIVEN	35	0.64	0.74
BOTH	35	0.64	0.72
AGRICULTURAL	32	0.56	0.34
FUTURE	32	0.56	0.75
GENERAL	31	0.53	0.86
TERM	31	0.53	0.84

*= words classified as academic only when nouns

Let us now turn our attention back to the F-SCUSSE, but once again from the perspective of it being divided into two sub-corpora. This approach was already utilized when gathering the data for figure 4, although now the focus is on individual academic words. As previously mentioned in that context, the token sizes of the two sub-corpora are slightly imbalanced (11,000 tokens in contrast to 52,000 tokens). Therefore when we contrast the most frequently used AWL items across the two levels, this imbalance is shown in table 10 by the great differences in word frequencies. For instance, *technology* appears 11 times in first year students' essays while it appears 115 times in the writings of third year students. This is why the numbers are also presented as occurrences per 1,000 words. There are not many words with a frequency above 1.00 per 1,000 words, but *role*, *example*, *need*, and especially *important* attract attention. If the sample size was greater, it might be expected that more words attract attention. The sample size is larger for the third year students, and thus words with a higher frequency than 1.00 per 1,000 words appear more regularly. This is certainly no rule, and the numbers are products of several causes, but an increase in sample size can however be expected to function in such a way.

Table 10. Most common academic words (AWL) in the F-SCUSSE according to level

1st year students			3rd year students		
Word (AWL)	**Freq.**	**Per 1,000**	**Word (AWL)**	**Freq.**	**Per 1,000**
ROLE	13	1.19	TECHNOLOGY	114	2.20
TECHNOLOGY	10	0.91	MEDIA	96	1.85
MENTAL	8	0.73	NUCLEAR	56	1.08
PHYSICAL	7	0.64	AUTHORITIES	34	0.65
CONTRIBUTES	6	0.55	PHYSICAL	33	0.64
AFFECT	6	0.55	ENVIRONMENT	29	0.56
TOPIC	5	0.46	ENERGY	26	0.50
STRESS	5	0.46	RESEARCH	23	0.54
INSTANCE	5	0.46	POSITIVE	21	0.40
DEPRESSION	5	0.46	RELAX	20	0.39
NORMAL	4	0.37	JOBS	20	0.39
MAJOR	4	0.37	COMPUTERS	20	0.39
GENDER	4	0.37	INSTANCE	19	0.37
BENEFITS	4	0.37	TOPIC	17	0.33
AFFECTS	4	0.37	CREATE	16	0.31
MEDIA	3	0.27	NEGATIVE	15	0.29
MAINTAIN	3	0.27	FOUND	15	0.29
ITEMS	3	0.27	ISSUE	14	0.27
ECONOMY	3	0.27	CULTURES	14	0.27
COUPLE	3	0.27	AID	14	0.27

Turning to the AVL, we rediscover the vocabulary variation that separate the lists from each other. The data in table 11 shows several more items with a frequency above 1.00 per 1,000 words than what was showed in table 10. It is noteworthy that the top twenty academic words, both in terms of the AWL and the AVL, are more similar to the NS learner corpus for the third year students, than they are for the first year students. In fact, many non-topic-tied words seem to follow a pattern, where they appear occasionally in the writings of NNS first year students, quite frequently in the writings of NNS third year students, and most frequently in the writings of NS students. Examples of this phenomenon include *however*, *use* and *issue*. This pattern seems to be even more prominent when we look at it from the opposite direction, from the most frequent in younger NNS writing to the least frequent in NS writing. Such tendencies appear in the cases *need*, *example*, *both*, and *important*. As to the topic-tied words, we can observe interesting distributions. In the previous section on general frequencies we discovered that the most frequent words used by NNS learners tend to be general words, even when function words were ignored (cf. table 2.2 and table 5), in contrast to NS learners where topics emerged. Unsurprisingly, given that we know what the academic word lists include, they allow us to identify the topic-

tied words. In other words, the academic vocabulary lists make words such as *technology*, *future*, *language* and *climate* become visible. Thus, even the topics of the NNS learner become clear.

Table 11. Most common academic words (AVL) in the F-SCUSSE according to level

1st year students			**3rd year students**		
Word (AVL)	**Freq.**	**Per 1,000**	**Word (AVL)**	**Freq.**	**Per 1,000**
IMPORTANT	45	4.11	HOWEVER	151	2.91
NEED*	29	2.65	SOCIAL	121	2.33
EXAMPLE	15	1.37	TECHNOLOGY	114	2.20
ROLE	13	1.19	EXAMPLE	108	2.08
TECHNOLOGY	10	0.91	IMPORTANT	104	2.00
FUTURE	9	0.82	NEED*	80	1.54
MODEL	8	073	CHANGE*	79	1.52
MENTAL	8	0.73	QUALITY	70	1.35
BOTH	8	0.73	FUTURE	65	1.25
MEAN*	6	0.55	CLIMATE	64	1.23
AFFECT	6	0.55	USE*	60	1.16
WHOLE	5	0.46	SOCIETY	45	0.87
STRESS	5	0.46	COMMON	37	0.71
MODERN	5	0.46	BOTH	32	0.62
LANGUAGE	5	0.46	SCIENCE	30	0.58
KNOWN	5	0.46	EXPERIENCE	29	0.56
INSTANCE	5	0.46	ENVIRONMENT	29	0.56
HUMAN	5	0.46	INFORMATION	28	0.54
DEPRESSION	5	0.46	NATURAL	27	0.52
WORKING	4	0.37	MEAN*	27	0.52

*= words classified as academic only when nouns

When we work with corpus data we sometimes notice certain features not specially sought for. Such was the case when I analysed the ordinal frequency lists of the academic word lists' type and token coverages in the F-SCUSSE. The comparisons revealed three groups of words: matches between an academic word list and the learner corpus, non-academic vocabulary from the NNS learners, and numerous academic words not used by NNS learners. To investigate those final cases, I compiled a list of items that had not been used in the NNS learner essays. There were several hundred ones to choose from, which meant that the selection had to be structured. Two criteria formed the basis for this structural approach. First, the list was examined alphabetically and approximately every 500th item was chosen for closer inspection. Secondly, the decision was made that at least one word family member must appear in the top 5,000 most frequent words of the BNC for it to be included, to avoid very rare lexical items.

Therefore, if an item fulfilled the first criterion but not the second, the next item in the list was selected instead. For instance, the word family of *sustain* includes the word family members *sustain*, *sustains*, *sustained*, *sustaining* and *sustainability* in the AWL. For this word family, the word *sustained* appears as the 4,281st most frequent item in the BNC. The other word family members are not as high in frequency, but it still means that the item may be included in the list. An example of an item not considered suitable for inclusion is the item *denote*, whose word family member *denote* appears as the 16,708th most frequent item in the BNC. Thus, table 12 presents the results of the selection of the academic word families in the AWL not represented in the learner essays in the F-SCUSSE. The verb forms of each word family function as representatives of the word families in the examples. Interestingly, many of these word families are represented in the LOCNESS A-levels essays. Considering what has been found about lexical richness in NS writing (cf. Paquot & Granger, 2009a), it could have been expected for varied word families to occur more regularly in NS student essays than in NNS learner writing, but the fact that they occur in 10 out of 12 cases was above expectations.

Table 12. Word families (AWL) not represented in the F-SCUSSE

Word family	Represented in the LOCNESS A-levels essays? (yes/no)
ACKNOWLEDGE	Yes
EXCLUDE	Yes
INDICATE	Yes
INTERPRET	Yes
PROCEED	Yes
RESTORE	No
REVEAL	Yes
SELECT	Yes
SUMMARIZE	No
SUSTAIN	Yes
SYMBOLIZE	Yes
VALIDATE	Yes

The next step was to examine how frequently these word families were used in NS student writing, given the coincidence that so many of them do occur. When performing a manual search it was found that members of the word families are utilized primarily once or twice in the NS corpus. This gives grounds to two observations: Firstly, the language of NS learners is more varied, and secondly, even though members of word families are used, they are not used particularly frequently. Most word families in the

selection are utilized in the NS corpus, except for the word families of *restore*, which includes *restore*, *restores*, *restored*, *restoring*, and *restoration,* and *summary*, which includes *summary*, *summarise*, *summarises*, *summarised*, *summarising*, *summaries*, *summarisation*, and *summarisations*.

It was previously established that the AVL does not contain word families, but instead it has been constructed on the basis of the idea of core academic words. Therefore, it is not possible to perform a direct comparison between the two lists per se. Nevertheless, the approach used for the compilation of table 12 was utilized when the results for table 13 were collected. In other words, table 13 presents a selection of some of the core academic words in the AVL, as well as their word family members, which are not represented in the writings by the contributors of the F-SCUSSE. For instance, when determining whether the item *calculate* was represented either in the F-SCUSSE or in the LOCNESS, all of its word family members were examined as well, for instance, *calculated*, *calculating*, and *calculation*, i.e. exactly like when the AWL's word families were checked.

Table 13. Core academic words (AVL) not represented in the F-SCUSSE

Core academic word	Represented in the LOCNESS A-levels essays? (yes/no)
ASSESS	Yes
ASSOCIATE	Yes
CALCULATE	Yes
CONSTITUTE	Yes
DETERMINE	Yes
EMERGE	Yes
IDENTIFY	Yes
ILLUSTRATE	Yes
OBTAIN	Yes
VALIDATE	Yes
VARY	Yes
VERIFY	No

The pattern that emerges is similar to the pattern of table 12, even though the core academic words and the word family members belonging to them are different. This time too, the words were chosen from the cases that the restrictions had made possible, and it is merely a coincidence that so many appear in the writings that constitute the LOCNESS A-levels essays. As in table 12, the fact that the NS learners have used the items in their writings does not mean that they are particularly high in frequency. Instead, I would propose that they are signs of lexical richness and individual variation.

6. Discussion

The present study has been a comparison of learners' vocabulary use according to several classifications. There has been an internal comparison of both general and academic words within two learner corpora, but also external comparisons (Cobb & Hurst, 2015: 189), in this case with the BNC as a reference corpus. In addition, the learner corpus in focus, namely the F-SCUSSE, has been divided into sub-corpora according to the level of the learners. The part of the LOCNESS I have focused on contains but one level, which means that no further division was required. Furthermore, this has been a comparison of the practical implementations of two academic word lists, namely the AWL and the AVL.

As previously mentioned, it has been suggested that academic vocabulary knowledge has positive effects throughout the complete educational system (cf. Gardner & Davies, 2013: 305; Zwiers, 2008). Throughout the present study, it has thus been assumed that the vocabulary of academic word lists is not only useful for younger students in their essay writing, but also necessary if learners are to succeed in their potential university studies. When in university, the Anglophone research tradition has meant that English is in a position above other languages, which subsequently manifests its status as the most influential second language in Finland. Moreover, the usefulness of the academic vocabulary cannot be related to English exclusively but also to the learners' presumed L1, Swedish. In other words, the writing conventions of one language, those that exceed the lexis in particular, may to some extent be implemented in another. All things considered, the perceived importance of English as rated by Finns themselves (cf. Leppänen et al., 2011) , in combination with globalisation and an ever increasing demand for a high English proficiency, have put English in a favourable situation in Finland.

In the present study I have treated the NS varieties in the LOCNESS and the BNC as norms, without any genuine justification. It should therefore be explained why a norm has been implemented and utilized in the analysis, and why we classify norms as *expert writing* (cf. Granger & Paquot, 2009a: 210). The categorisation of some varieties as norms and others as striving for the norm may appear controversial at first, especially as it neglects the idea that the learner language may be a variety in its own right (Callies, 2015a: 49). Thus, to draw any conclusions from the observation that the word *like* appears twice as often in the NNS corpus as in the NS corpus might be valuable, although simultaneously questionable, information. However, not

everybody agrees that NNS learner varieties should be treated as equals to NS varieties. Ai and Lu even suggest that NS varieties are not given enough recognition as a norm when they propose that a NS "baseline appears to be a rather neglected dimension" (2013: 249) when studying NNS proficiency in the target language. Even as the case may be that a NNS variety is its own rightful variety, it is one that undergoes a lot of development. Most often, this development is in the preferred direction of better overall proficiency in the target language, which is subsequently why there is a need for a norm variety in the first place, i.e. an ultimate goal to achieve. Furthermore, as a variety undergoes a lot of development it may turn out to be difficult to pinpoint its characteristics. It seems that the preferred approach is to combine the two perspectives. With this in mind, we must not treat NNS learner language as inferior, even if a NS baseline in LCR is required because it shows what the NNS learners should strive for.

In connection to the previous section, the premise has also been that academic vocabulary knowledge functions as a gateway to native-like fluency in writing. The term *fluency* is closely connected to the term *native speaker*, which in itself is a troublesome concept with ill-defined boundaries. Understandably, this notion might seem problematic but let us take a closer look at the term *fluency*. Fluency development is a term Nation (2001: 2-3; 127-129) commonly discusses, and which is categorised as a member of the four strands of language learning. According to the principle the three other strands include meaning focused input, meaning focused output and language focused learning. This would make fluency development into an essential part of classroom teaching. Previously, we have discussed some of the goals of EFL teaching, where native-like fluency was considered a component of a long-term goal. However, what does native-like fluency incorporate in practice? Is fluency being able to speak quickly? Alternatively, is it being able to write unhindered? Arguably, the answer to these questions could be yes, perhaps accompanied by a commentary. Nevertheless, we could argue that fluency implies being able to choose your words carefully according to the context. Thus, in this particular context of argumentative essays, a prominent use of academic vocabulary could indeed suggest a higher degree of the fluency that is typically associated with native speakers.

At this stage it should be mentioned that there were difficulties involved in the comparison of the results to previous studies. Even as several previous LCR studies were found, those that concerned intermediate learners were scarce in number. Instead, most of them dealt with advanced learners at university level. Similarly, there

are further imbalances in the context of learner corpora, where, for instance, more written than spoken corpora and more general than specific corpora are available. This simply relates to the availability of advanced learner materials in contrast to intermediate learner materials (Gilquin, 2015: 28). Of course, we could argue that for some learners the variation in level between an advanced learner and an intermediate learner might not even be that substantial. Nonetheless, keeping the two levels separate creates structure, as there might be considerable age gaps, which consequently leads to the conclusion that more data on younger learners is needed. The process of collecting data from younger learners might be more demanding and not as convenient, but it is well worth the effort.

6.1 Frequencies

Frequencies of words have formed the basis for substantial parts of the present study. This approach of analysing frequency data has been for long, and is likely to continue to be, a common methodological procedure in the field of LCR. In fact, frequencies of words and word combinations have become increasingly relevant, as a result of the recognition LCR has gained in the field of language learning. Leech states that "frequency information remains a highly valuable resource for input to language learning materials and testing" (2011: 27). This all relates to the idea of how important it is to master the vocabulary whenever we try to learn a new language (Cobb & Hurst, 2015: 185).

For the present study, the idea was to divide my approach in the field into an initial examination of frequencies of general vocabulary before advancing to the analysis of academic vocabulary. One of the reasons to why such an approach was considered appropriate was that the general vocabulary frequencies might show cases of limitations that could be improved with the use of academic vocabulary. Furthermore, the advantage of this type of procedure is that it gives the researcher an opportunity to discover the results black on white, whether they are sought for or not. In fact, much of the data collected from the analysis of the NNS learner materials was not to be sought for at first but instead it appeared as a by-product of interpreting the results.

In contrast to Ringbom's (1998) focus on advanced learners at university level, the focal point was to investigate the academic vocabulary use of even younger learners. These younger learners have been categorised as intermediate learners, even

though the difference in level in contrast to advanced learners might not be major, or at least not for the higher level students in upper secondary school. There is indeed the issue of chronological variability between Ringbom's study and mine, but the results can still be considered comparable as the topics are related.

Ringbom (1998; 1999) noticed that especially two words appear to stand out in the writings of advanced learners: *people* and *thing*/*-s*. The present study provides new information that confirms his observations at least to some extent. In fact, there seem to be many additional words in the F-SCUSSE that could be described in similar ways, for instance, *good*, *nice*, and *think*. Moreover, comparisons made between the F-SCUSSE and the LOCNESS offer additional information on vagueness. In the LOCNESS, the word *people* appears almost as frequently as among Finland-Swedish students in the F-SCUSSE. This would signify that its use does not only depend on proficiency level, but that it could be related to linguistic maturity, which ties in with the phenomenon of lexical teddy bears (cf. Hasselgren, 1994; Leech, 2011: 14). Normally, the use of vague words could be seen as a sign of a vague and stereotyped language but "concrete evidence of exactly what constitutes this vagueness has been hard to come by" (Ringbom, 1998: 49). Then again, it could be that the English language lacks options for replacing the word *people*, without rephrasing the text too much. It may not be the ideal approach to speak of *overused* and *underused* lexical items (cf. Leech, 1998: xix-xx, in Granger, 2009: 22) when describing occurrences in learner corpora. The very definition of what makes a word over- or underused is problematic, but for lack of more satisfactory terms, they are utilized in the following examples. The data suggest that the collocation *some people* may in fact be overused in the F-SCUSSE, when compared to simply using *some*. The same is true for the collocation *many people*, instead of just referring to individuals as *many* or the collocation *other people*, instead of *others*. In fact, *other people* did not occur even once in the writings of NS students, which seems to confirm such an observation.

> EFL learners significantly underuse the majority of 'academic verbs' […] when learners use academic verbs, they tend to restrict themselves to a very limited range of patterns, which contrasts sharply with the rich patterning that characterizes expert writing.
>
> (Granger & Paquot, 2009a: 210)

The results show that the NNS learners tend to use mostly "conversational verbs" (Granger & Paquot, 2009a: 210) in their writings, such as *think*, *like*, *make* and *want* as opposed to verbs considered part of the academic vocabulary (cf. Doró, 2015: 71-72). As previously discussed, the function of the latter is for example to organise, exemplify and summarise in a text (cf. Granger & Paquot, 2009: 98). Verbs in general tend not to occur very commonly among the most frequent words in either the NNS corpus or the NS corpus, even when the stop list is utilized, which could suggest a high degree of personal stylisation. What I mean is that learners might use the more varied repertoire of everyday verbs, such as *read*, *break*, *enjoy* and *remember* instead of using the limited conventional verbs typically associated with academic writing. Obviously, the perception of the argumentative essay in terms of its content is that it should not be as restrictive as for academic writing, but the end product can easily appear too subjective if the writing is not controlled in any way. Furthermore, Finland-Swedish intermediate learners also tend to use the preposition *of* modestly in contrast to the writers and speakers of the two NS reference corpora. This probably stems from an extensive use of the *'s*-genitive and thus an underuse of the *of*-genitive, which in turn could be explained through L1 interference. In Swedish, the genitive case is marked almost exclusively by a final *s*.

Cobb and Hurst (2015: 188) propose that personal pronouns tend to be used extensively in EFL learner texts. The results of the present study show that the personal pronouns *you* and *I* are high-frequency items in the F-SCUSSE, whereas neither of the items occurs among the top 15 most frequent words in the LOCNESS. *I* does however appear as the 12th most frequent item in the BNC, which confirms that the word is used across many genres. The use of these pronouns in argumentative essays is part of the reason why teachers describe learner essays as "overly personal and speech-like in style" (ibid.). There are two established hypotheses that could possibly explain this wide use of the nominative case of pronouns. The differing use of pronouns in Swedish and in English could be one explanation, i.e. L1 interference once again. The generic pronouns are used differently in Swedish and in English, which would for instance explain why *you* occurs so frequently. Another factor relates to the influence from informal spoken language. The latter does make sense, especially if we consider the contexts in which the learners encounter English outside of the classroom. Moreover, a third explanation, which also seems likely, could be derived from the learners' inability to simply distance themselves from their own texts, as they have not

yet progressed from the sequential narrative to the logical structuring of explanation (cf. Zwiers, 2008: 196).

In terms of academic vocabulary the results show tendencies for similarities between NNS writers and NS writers to some extent, even though the differences that set them apart are enough to justify a clear division between the two varieties of writing. The higher level NNS learners in the F-SCUSSE and the NS students in the LOCNESS are especially similar in their use of academic vocabulary. Of course, we must not forget that lower level students have not had the opportunities to practise their argumentative writing to the same extent as the higher level students. Similarly, to develop on Lindgren's (2015) conclusions the difference in level between advanced learners and intermediate learners make for an interesting observation. In her BATMAT corpus the AWL had a mean token coverage of 6.9 %, in contrast to the LOCNESS, 4.9 %, and the F-SCUSSE, 3.3 %. The apparent gap between academic texts written in university and argumentative texts written by EFL learners in lower education demonstrates just how widely used academic vocabulary is in higher education. Furthermore, Lindgren's study included an element that would not feature in the present study, namely grading, which was not possible to incorporate, as the materials shared with me were not corrected or graded. Nevertheless, her findings suggested that there was no connection between a frequent use of academic vocabulary and a higher grade, which in itself makes for an intriguing discussion.

Many academic words, or at least individual lemmas from those academic vocabulary families, occur in the LOCNESS but not in the F-SCUSSE. As previously stated however, many of these words do not occur very commonly. They may occur only once or twice in the writings of NS learners, which instead points to a greater lexical richness, as well as a larger degree of individual variation.

I would like to highlight a few word families included in either the AVL or the AWL or in both, but not represented in any of the student texts. Similarly to when Schmitt and Schmitt (2005: iv) suggest that there are words learners should already know, these word families would probably belong to the same category. These include families such as *identify* and *illustrate*, which appear in both academic word lists, and *remove* and *restore*, which appear in one of the lists. My guess is that many of the learners recognise these word families, even though they have not made use of them. Perhaps some students simply did not find it necessary to make use of such words, but it is definitely a coincidence that not a single NNS learner had incorporated

them in their texts. In the field of LCR however, we cannot believe in coincidences, as we can only examine the vocabulary that has been used. Therefore, it makes for an interesting observation that the word families not appear in either learner corpora.

In what has been a study of vocabulary frequencies, it would seem suitable to mention Sinclair (1991; 2004) and his contributions to linguistics that are applicable in this context. He introduced two concepts in terms of text production*: the idiom principle* and *lexical grammar*. The former implies that words do not appear in isolation, whereas lexical grammar refers to the relationship between grammar and vocabulary, including for instance the role of chunks, i.e. groups of words, in grammar acquisition. The opposite of the idiom principle would thereby be the open-choice principle, also coined by Sinclair (1991), which states that every single word in a sentence is a deliberate choice. Corpus data has established that words tend to occur in chunks rather than in isolation (Ädel, 2015: 413), which would lend credibility to the idiom principle. Leech (2011: 15) too questions whether frequencies of word combinations are more important than frequencies of individual words, if the matter is considered from a learner perspective.

The words investigated in the present study have only been examined in an isolated setting, except in terms of the previously discussed *people* and its most common co-occurrences. In that particular case, it was observed that what separated the use of *people* actually relates to the choices of collocators. Therefore, the differing uses of collocations between NS and NNS learners confirm that they may reveal valuable information about the item of interest. We can thus conclude that frequencies of co-occurrences, or collocations, could form the next logical step after having examined frequencies of words in isolated settings.

6.2 Academic Word Lists

It is reasonable to question whether it is justifiable to study the coverage of academic word lists in texts written by students who have probably, although not certainly, never read an academic text. As previously stated, academic language is a type of language most of us encounter in its full form only in higher education. To repeat what was discussed previously, academic language incorporates considerably much more than simply the lexis of a wide range of disciplines in higher education. It incorporates a way of thinking and a way of structuring text accordingly (cf. Zwiers, 2008). It has also been determined that the educational system is a cumulative system, where lower level

learners are prepared for a higher level. For instance, high school learners are prepared for upper secondary school, and students in upper secondary school are prepared for university. In the context of upper secondary school learners, I would then propose that there is genuine interest involved to investigate to which degree their language corresponds to academic standards. This is especially true now that English has proven to be highly influential, some would say unavoidable, in higher education. Many courses in Finnish universities are in fact in English, and the course materials are rarely in any other language than English.

Another justifiable cause relates to the content of argumentative writing. The proposed token coverage regarding the genre of fiction was given for both academic word lists by their respective creators. This number was 1.4 % for the AWL (Coxhead, 2000: 213) and 3.4 % for the AVL (Gardner & Davies, 2013: 323). In contrast, the AWL and the AVL had a token coverage of 3.3 % and 5.7 % respectively in the F-SCUSSE, and 4.9 % and 7.5 % in the LOCNESS A-levels essays. Thus, the fact that the token coverage for the learner data exceeds the lists' token coverage of fiction suggests that the genre of argumentative essays constitutes a middle ground between fiction and academic writing. Students writing argumentatively cannot use stylistic markers typically associated with fiction too extensively, such as rich imagery, when they try to get their message across. In other words, students recognise that they need to alter their language in argumentative writing, whether they be native speakers or intermediate learners.

Whether it is favourable to extract lemmas or word families out of academic corpora to compile academic word lists is difficult to determine based on the experiences of implementing them only once. What can be commented on is that there seems to be a considerable share of subjectivity incorporated in the choice of these lexical items, as done by the researchers. Similarly, we find that the implementation and utilization of academic vocabulary lists follow similar patterns, for instance by educators and course book designers. For example, when Schmitt and Schmitt (2005) designed their course book for mastering the academic vocabulary, they based it on the AWL. As previously mentioned, the AWL consists of 570 words families, but they chose to incorporate only 504 word families, leaving out 66. The course book targeted at advanced learners mixes the theory by Nation (1990) with Coxhead's word list. However, they assumed that the students "should already know" (Schmitt & Schmitt, 2005: vi) certain common words, such as *area*, *require* and *similar*. The decision is

somewhat questionable, and it demonstrates the importance of the choices made by those working within the linguistic field.

The AWL has been designed to include a wide range of academic vocabulary but not to cover non-academic texts substantially. Nevertheless, the AVL succeeds better in coverage not only for academic texts but also for the argumentative writing in the NNS essays, even if the two lists are equal in terms of length. This distinction between the two lists ought to be highlighted, and could function as a major deciding factor when choosing a preferred version to work with or implement in the curriculum. Continuing on this topic, Gardner and Davies do not agree with the notion that academic vocabulary should follow after a high frequency list, for instance after the 2,000 first words in the GSL as is the case with the AWL. This means that they are not "concerned with the fact that many core academic words may appear in the highest frequency lists of [any large corpora]" (2013: 310). They even expect that several of the words included in their list would occur towards the top of an ordinal frequency ranking. As a result, the AVL tends to include more items that could serve a vast range of purposes in an academic text. These words include for instance, *however* (adverb), *therefore* (adverb), *study* (noun) and *research* (noun). These are all words that occur very frequently in academic writing, even though they might not belong to the most semantically heavy words or relate to a specific discipline. The purpose of these words is rather to link sections and arguments, and to make the text progress smoothly. In contrast, the words included in the AWL appear more meaning bearing. However, from a learning and teaching perspective, the AWL might still be preferred, as it has been constructed from just 570 words families, whereas the AVL has been constructed from considerably more word families. Thus, the learning burden of such a list as the AVL becomes inevitably greater.

An additional aim of the present study has been to determine how viable the two academic word lists are when used for research purposes. The format of the AVL is more complex than the format of the AWL, when instead of word families it utilizes lemmas. This means that when used correctly, the AVL would result in more precise conclusions (cf. Hartshorn & Hart, 2016: 84). However, when working with the AVL it would definitely be advisable to use some form of POS-tagging. Otherwise, if the AVL is converted into word families the results may show signs of overinflated data. Both Hernandez (2017) and Newman (2017) encountered almost identical problems and ended up with similarly skewed results. The margins of error in their

results, as in mine, are minor but nevertheless the issue should be acknowledged. As previously mentioned, this becomes especially evident for words such as *use* and *need*. In conclusion, both academic word lists proved to be applicable and viable. However, anyone interested in using them should consider his or her purposes carefully before selecting which one to work with.

6.3 Implications for Teaching

> Because of researchers' tendency to collect data from learners who are easy to reach, we also notice a predominance of learner corpora representing relatively advanced university students […] whereas beginners and young learners are less often represented.
>
> (Gilquin, 2015: 28)

For the present study, the focus has been on Finland-Swedish learners in upper-secondary school in Finland. The learners are younger than advanced university students, which is important to keep in mind in relation to the comments made by Gilquin. Similar to many studies of learner corpora before this one, there has been an underlying intention to offer practical teaching applications based on the results. This is not only due to a personal interest in classroom techniques, but also due to a responsibility felt towards the many teachers and students who were kind enough to give me access to their students' materials. In a sense, the F-SCUSSE can be described as a local corpus, for the reasons mentioned above. A local learner corpus, which presumably is similar to an IPU, or a learner corpus for immediate pedagogical use, invites "teachers and students alike into the field of LCR by making them both providers and beneficiaries" (Gilquin, 2015: 29). Nevertheless, a cautious approach is advised because we do not wish to draw any conclusions based on mere speculation. In addition, the F-SCUSSE represents learner English of the Finland-Swedish students only, and not for instance Finnish or Swedish students.

Frequency results are indeed very valuable, but we must also remember that they are mere results that give the impressions that item X appears more frequently than item Y. Similarly, Chambers (2015: 462) argues that there "is thus a considerable need for research into the integration of learner corpus data in language learning and teaching", whereas Gries (2015: 175) suggests that "in a single test, the probability of

erroneously assuming that a finding is significant is typically 5 %", which further reinforces the notion of carefulness. Nesselhauf (2004) agrees that frequencies can be troublesome, especially when the researcher is to offer implications for teaching:

> Teaching recommendations exclusively based on the criterion of frequency of deviation in non-native speaker usage seems [sic] similarly misguided
>
> (Nesselhauf, 2004: 119)

Myles (2015: 313) argues that essay writing is not a "reliable window into [learners'] underpinning linguistic system". It could however be argued that if essay writing is not a valuable resource for studying the linguistic system of learners, then what is? A move away from the study of the relatively free form of essay writing probably implies a move towards a more controlled research setting, whether we refer to spoken or written language. If so, we would most likely find ourselves in a similar situation to some decades ago when "the data used was rather artificial" (Granger et al., 2015: 1). Of course, it may be wise to set limitations in terms of for instance genre. Nevertheless, our intention should always be to seek out as naturalistic data as possible, even if it may lead to compromises and issues of control.

When comparing levels, or first year students to third year students, it does in a way represent a pseudo-longitudinal approach, or as Meunier (2015: 381) describes it, a "comparison of cross-sectional studies of different groups of learners at different developmental stages". Instead of studying the development of language in a traditional longitudinal manner, where time may prove to be a challenging variable for the corpus design, a pseudo-longitudinal approach may offer similar results. This is an effective method, even though the learners represent two separate groups of people. Of course, this also implies that the groups who represent the development in a pseudo-longitudinal study share several characteristics, which the first year students and the third students in the F-SCUSSE do. The characteristics shared include for instance, L1, education and a similar goal: to eventually pass the matriculation exam. However, one issue would be the difference in sample size, where as previously mentioned, the younger students' writings merely account for roughly a sixth of the F-SCUSSE, whereas the older students' writings account for the rest.

There would be reason to focus even more on genre-specific types of writing, especially in the sense of argumentative writing. This way learners would

begin to acknowledge the specific stylistic requirements both in terms of its form and its content. News articles, letters to the editors and popular reports are just examples of some of the types of texts the learners could familiarize themselves with. These types of texts are presumably already used by several teachers, and it may be easier said than done to increase the amount of learner exposure to such objects. Furthermore, as Hinkel (2003: 85) also points out, the so called "rare words" are troublesome because they are difficult to encounter in everyday situations.

> Another factor that greatly complicates the learning of L2 academic vocabulary is that it is not the common words that create the greatest difficulties in reading and writing, but the relatively rare words that actually represent the largest number of words used even in basic academic texts.

However, the initial approach should always be for students to familiarize themselves with, and acquire the common words, before progressing to the rare ones.

What has also been discovered is that learners should perhaps be made more aware of the distinction between written and spoken language. Again this is problematic because spoken language covers a substantial part of the type of English that young learners encounter on a daily basis. There is one practical task that could function as a tool to aid in the process of becoming more aware of the distinction between written and spoken language. This would include using the hands-on approach of the materials gathered by the teacher, which would then inform learners of which words to avoid, and which to practice. Previous studies have observed that learners tend to be receptive to and appreciate the data gathered from their own writings (Cheng et al., 2003), which could then function as motivation for improvement.

The collocations of the word *people* were discussed previously. In language learning the focus is sometimes on individual words, and certain expressions, and not so much on the variety of combinations at the learners' disposal. Thus, collocations could be a focus that is given more thought in the classroom. With that in mind, collocations would follow the same recommended principles as regular lexical items, with the most frequent ones being the most important for EFL learners (Nation, 2001: 322-323).

7. Conclusion

At the outset of the present study, I aimed to answer four particular questions. These questions functioned as tools for studying the language that constitutes Finland-Swedish intermediate EGAP, or argumentative writing. However, just like Hasselgård (2009: 138) suggests, "the exploration of learner corpora can lead to insights that were not even sought for at the outset of the investigation". Thus, the following observations can be noticed from the results.

What is the coverage of the Academic Word List, AWL, and the Academic Vocabulary List, AVL, in the Finland-Swedish Corpus of Upper Secondary School English, F-SCUSSE, in contrast to the A-levels essays of the Louvain Corpus of Native English Essays, LOCNESS?

The AWL contributes for a token coverage of 3.3 % in the F-SCUSSE and 4.9 % in the LOCNESS, whereas it covers 11.3 % of types in the F-SCUSSE and 12.7 % in the LOCNESS. The AVL covers 5.7 % of tokens in the F-SCUSSE and 7.5 % of tokens in the LOCNESS. It has a type coverage of 11.7 % in the F-SCUSSE and 13.4 % in the LOCNESS. All numbers suggest that NS students in the LOCNESS use academic vocabulary to a larger extent than the NNS students in the F-SCUSSE. This seems to hold true both in terms of frequency and in terms of variation.

What academic lexical items, as suggested by the AWL and the AVL, are the most frequent in the F-SCUSSE?

The five most frequently occurring items from the AWL include *technology*, *media*, *nuclear*, *physical* and *authorities*. According to the AVL, the same procedure includes the items *however*, *important*, *technology*, *social* and *example*. By and large, the words suggested by the AWL tend to be more topical than those suggested by the AVL, whereas the words incorporated in the AVL can be said to serve several functions, for instance to create links in the text. This is an important distinctive factor to consider when working with one of the lists or both. Nonetheless, the two lists share numerous cases.

In terms of the most frequently occurring general items, the NNS writers of the materials that constitute the F-SCUSSE show no substantial dissimilarities in the use of function words except in a few cases. However, when a stop list is applied the most frequently occurring general items suggest a restricted variation in word choices. Vague words, linked to the phenomenon of lexical teddy bears, such as *people*, *think*,

like and *good* appear more prominently in the essays of the F-SCUSSE than in the LOCNESS. Such words could easily be substituted for synonyms to give the texts more depth, which in turn would increase the perceived EFL proficiency, and thus fluency. Nevertheless, we do notice that this lexical development increases with the level of the learner, which is an encouraging sign.

How does the use of academic vocabulary vary between different levels of learners in the F-SCUSSE?

The results reveal tendencies that higher level NNS learners use a more topic-tied vocabulary than their lower level counterparts judging by the coverage results from the academic vocabulary lists. The AWL contributes for a type coverage of 8.1 % for 1st year students and 11.9 % for 3rd year students. The AWL's token coverage is 2.2 % for the former, and 3.5 % for the latter.

Similarly, the AVL covers 8.7 % of types in the writings of 1st year students and 12.1 % of types in the essays of 3rd year students. The token coverage is 3.5 % for the former and 6.2 % for the latter. Thus, the argument seems accurate irrespective of the academic vocabulary list we are referring to. One of the reasons why is probably because higher level learners have had more practice in writing argumentatively than the lower level learners. These tendencies should however be treated cautiously as the sample size of the lower level NNS learners in the F-SCUSSE remains relatively small at this stage. Therefore, more research is needed in the area.

How does NS written language compare to NNS written language, and to what extent does the level of the NNS learners influence this comparison?

NS written language appears to be more lexically varied, as the two learner corpora are approximately the same size in terms of tokens, but the LOCNESS A-levels students have used considerably more distinct words, 6,205, in contrast to NNS learners who have used 4,006 distinct words. The results also suggest that NS learners when writing argumentatively avoid vague words that could be linked to the phenomenon of lexical teddy bears, such as *like*, *things*, *lot* and *important*, to a higher degree than NNS learners do. Yet, some items suggest that there are surprising similarities between the two groups of learners, for example the use of *people*. However, a close examination of the word *people* revealed that the item differed in terms its most commonly occurring collocations.

The higher type and token coverage of both the AWL and the AVL in the NS learner corpus than in the NNS learner corpus give grounds for assuming that the former use more academic vocabulary. The difference is however not as great as it could have been expected to be. In addition, the fact that the third year students in the F-SCUSSE use a vocabulary closer to the native-speakers than the first year students do, could be a result of successful language learning in the classroom.

The present study has shown that a small and local corpus, such as the F-SCUSSE, can be used as a tool to examine linguistic phenomena in learners' language (cf. Doró, 2015: 72). As long as the corpus has been constructed following thorough criteria, the data may well be used for the development of teaching materials. This is especially realistic as the results have been contrasted with several reference corpora, one of which parallels the characteristics of the learner corpora in several ways. Nonetheless, the F-SCUSSE is still at an early stage. Thus, sticking to Sinclair's (1991: 18) principle that "a corpus should be as large as possible, and should keep on growing", and Cobb and Hurst's (2015: 205) argument that "[once] developed, corpora do not normally get thrown away", my wish is to either expand the F-SCUSSE or continue the study of the existing materials.

Potential future investigations would be advised to incorporate some form of automatic error-tagging and POS-tagging. These tools were not used in the present study and created both challenges and limitations. Rayson and Baron (2011), for instance, implemented the Variant Detector, VARD, which is a tool originally used in historical linguistics in their learner corpus with a 90 % success rate. In terms of POS-tagging, the *Constituent Likelihood Automatic Word-tagging system* (CLAWS) (Gardside & Smith, 1997) seems a viable option. It has consistently achieved 96-97 % accuracy in several genres, and a quick test-run suggested that it was able to handle problematic words in the F-SCUSSE well. As previously discussed, such problematic words included for instance homonyms that belong to several word classes.

The present study also leaves us with unanswered questions. It may seem paradoxical to mention that results in LCR may seem artificial, although not in the same sense that Granger (2015: 1) and Cook (1986: 13) uses the term. Instead, if words are taken out of context and then examined in an isolated setting, it becomes difficult to comment on anything besides their frequencies. In addition, the LOCNESS A-levels essays probably date back to the early or mid-90s, and as we all know, language

undergoes constant development. We can therefore question whether the F-SCUSSE would show even more similarities with a more recent NS learner corpus? Future studies could therefore include identifying academic phrases and expressions, as well as cross-referencing the vocabulary in texts with how they have been graded in a more recent corpus. Collocations would perhaps be the most valuable for pedagogical reasons, as the collocation results in the F-SCUSSE showed dissimilarities with the NS learners.

Acknowledgements

I wish to thank all the teachers who helped me in the process of data collection. I also wish to thank all the students. This study would not have been possible without either of your contributions.

Swedish Summary/Svensk sammanfattning

Akademisk vokabulär i finlandssvenska gymnasiestuderandes argumenterande engelska-essäer

Inledning

Kraven på studerandes kunskaper i engelska, både vad gäller deras muntliga och skriftliga textproduktion, höjs för varje dag. Elever bör behärska ett flertal genrer för att påvisa goda kunskaper i språket, och det här innefattar ett språkbruk anpassat till genrens form och innehåll. Gymnasiestuderanden i Finland övar sig i att skriva argumenterande texter på engelska, eftersom det i studentexamen ofta förekommer diskussionsfrågor. Den akademiska vokabulären framstår därför i det här sammanhanget som speciellt fördelaktig, med tanke på dess gränsöverskridande egenskaper.

Det behövs lämpliga metoder för att undersöka textföreteelser, eftersom det engelska språket består av drygt 250 miljoner ord (*OED*, 2017). Därför har korpuslingvistik, och speciellt dess delområde *learner corpus research*, LCR, etablerats som användbara metoder inom språkforskningen och dess undergren andraspråkforskningen. Metoden är inte helt problemfri, men den har sina fördelar i jämförelse med traditionella språktest i andraspråksforskningen (Olsson & Sylvén, 2017: 127). De metoder som vanligtvis associeras med LCR relaterar till studier av *naturligt språkbruk*, det vill säga språkbruk i relativt fri form. Essäskrivning utförs oftast med riktlinjer, men trots allt kan man argumentera att resultatet är nära naturligt språkbruk (jfr Granger, 2015: 1).

Sammanställningen och användningen av så kallade textkorpusar är inte ett nytt fenomen. Metoderna må ha digitaliserats, men redan när forskare räknade frekvenserna av enstaka ord manuellt betraktades mer frekventa ord som viktigare att lära sig (Leech, 2011: 8). Det här betyder inte att lärare och läroplansplanerare ska ignorera ord som uppvisar låg frekvens, men det kan vara ett hjälpmedel i själva planeringen av undervisningen. Genom att fokusera på en viss genre, eller en viss typ av vokabulär kan undervisningen effektiveras. Då kan ett akademiskt ordförråd, samt akademiska ordlistor framstå som speciellt användbara.

Syfte

Syftet med den här studien är att undersöka finlandssvenska gymnasiestuderandes allmänna och akademiska ordförråd på engelska. Ett annat underliggande syfte är att bidra till en större förståelse inom andraspråksinlärningen. Den specialiserade korpus, eller inlärarkorpus, som har sammanställts för den här studiens ändamål är F-SCUSSE (*Finland-Swedish Corpus of Upper Secondary School English)*, och innefattar cirka 260 engelska essäer skrivna av gymnasiestuderande från tre olika gymnasier i Svenskfinland. Målet är att evaluera elevernas språkbruk, och att sedan jämföra det materialet med två stycken så kallade akademiska ordlistor, AWL (*Academic Word List*, Coxhead, 2000), och AVL (*Academic Vocabulary List*, Gardner & Davies, 2013). Samma procedur kommer även att utföras på en del av inlärarkorpusen LOCNESS (*Louvain Corpus of Native English Essays*, CECL). Den del av LOCNESS som inkluderats innefattar texter skrivna av studerande på nivån *A-levels* i Storbritannien, vilket nivåmässigt motsvarar ungefär abiturienter i Finland. De här elevernas förstaspråk är engelska. I fortsättningen kommer den här referenskorpusen bara att benämnas som LOCNESS. De elever med engelska som modersmål kommer att benämnas som NS-elever (från engelskans *native speaker)* medan eleverna i F-SCUSSE faller inom benämningen NNS-elever (från engelskans *non-native speaker*). Ibland benämns NS-elever med hjälp av termen L1, som syftar på förstaspråk, och NNS-elever med termen L2, som syftar på andraspråk. Vidare kommer den allmänna korpusen BNC (*British National Corpus*) att fungera som en referenskorpus, eftersom den är tillräckligt stor (cirka 100 miljoner ord) för att representera den brittiska varieteten av det engelska språket. Den här processen kommer även att möjliggöra att på vägen till resultaten kunna dra slutsatser av de finlandssvenska studerandes generella engelska ordförråd. Utgångsläget med den här studien är alltså att försöka besvara följande frågor:

1. Hur mycket täcker AWL respektive AVL av den totala ordmängden i F-SCUSSE respektive *A-levels*-studerande i LOCNESS?
2. Vilka akademiska ord, enligt AWL och AVL, är mest vanligt förekommande i F-SCUSSE?
3. Hur skiljer sig användningen av akademisk vokabulär mellan elever på olika nivåer i F-SCUSSE?

4. Hur framstår NS-elevernas skriftspråk i jämförelse med NNS-elevernas skriftspråk, och hur påverkar NNS-elevernas nivå den här jämförelsen?

Bakgrund

LCR är i princip en blandning av teori och metoder från åtminstone fyra olika discipliner: korpuslingvistik, lingvistisk teori, språkdidaktik samt andraspråks-forskning (Prentice, 2017; Granger, 2009: 15). Sylviane Granger är en av frontfigurerna inom forskning om inlärarkorpusar och hennes resumé innefattar till exempel en av de största inlärarkorpusarna som är tillgänglig för allmänheten, ICLE (*International Corpus of Learner English*).

Två termer som ofta används inom korpuslingvistik är *typ* och *token*. Typ relaterar till distinkta ord i en text, medan termen token innefattar alla löpande ord i en text. I en simpel mening som ”Han är han” har vi alltså tre tokens men enbart två typer, eftersom *han* förekommer två gånger. F-SCUSSE består exempelvis av 63 000 tokens men ungefär 4 900 typer.

Tidigare forskning inom LCR har påvisat att studerande med engelska som andraspråk använder sig av vaga ord, personliga pronomen och konjunktioner i stor uträckning medan somliga funktionsord i stället förekommer mera sällan än hos studerande med engelska som förstaspråk (Ringbom, 1998; 1999). Lindgren (2015) å andra sidan, studerade universitetsstuderandes akademiska vokabulär med hjälp av sin BATMAT-korpus och fann att studerande på högre nivå använde sig av ett mer akademiskt språk än de på lägre nivå. Nation (2001: 9) påstår att elever bör bemästra ett oerhört stort antal ord för att komma upp i samma standard som personer med engelska som förstaspråk. Dock bör vi hålla i åtanke att även individer med engelska som förstaspråk kan göra språkliga fel (Granger, i Viana, 2007: 12). Därför bör vi förhålla oss kritiska även till korpusar som består av enbart modersmålstalare. Elever med engelska som andraspråk kan också sägas utgöra en egen varietet (Callies, 2015a: 29), och det kan därför vara skäl att ifrågasätta om en jämförelse med modersmålstalare är nödvändig överhuvudtaget.

Material och metod

De material som används i den här studien är F-SCUSSE, LOCNESS, BNC samt de två akademiska ordlistorna AWL och AVL. Den förstnämnda sammanställdes av mig

under hösten 2017. Här redogörs även för hur de engelska ordlistorna AWL och AVL har tillämpats, och de metoder som har använts inom ramen för studien presenteras i korthet.

F-SCUSSE representerar, som tidigare nämnts, finlandssvenska gymnasiestuderandens skriftliga engelska språkbruk i argumenterande texter. Fyra stycken lärare gick med på att dela med sig av deras elevers engelska-essäer, förutsatt att eleverna själva, eller, ifall att de var minderåriga, deras vårdnadshavare gav sitt samtycke. Inlärarkorpusen innefattar totalt 263 argumenterande essäer, vilka utgör cirka 63 000 ord eller tokens, som har skrivits som del av gymnasiets kurser som övning inför studentexamen. De studerande som är representerade kan delas in i två nivåer: förstaårsstuderande samt tredjeårs-studerande. Essäerna på engelska hade blivit skrivna på dator, vilket möjliggjorde en smidig överföring från lärarna till mig. Eleverna hade dock inte använt sig av några hjälpmedel, som till exempel stavningskontroll, och essäerna hade inte blivit rättade eller betygsatta. På grund av det här rättades stavningsfelen manuellt, eftersom programvaran som användes, WordSmith Tools, kan feltolka felstavningar. De 263 essäerna sammanställdes till slut till en enda stor fil, vilken i sin tur delades in i två filer med avseende på nivå. Det genomfördes inte heller någon ordklasstaggning, vilket i efterhand gav upphov till en del överkomliga problem.

Den programvara som använts för frekvens- och konkordans-resultaten samt jämförelserna är WordSmith Tools, version 7. Den här program-varan används för att finna mönster i olika typer av texter (Scott, 2018). Främst användes funktionerna ”*Concord*” och ”*WordList*”. Concord används för att finna ett ord i sin ursprungskontext samt för att få värdet frekvens/1000 ord, medan *WordList* är den funktion som listar ord enligt frekvens. Det kan vara skäl att använda sig av en så kallad *stop list*, det vill säga en stopplista, om man vill filtrera bort exempelvis stoppord såsom *the*, *a* och *an*. Den stopplista som användes ingick i programvaran, och inkluderade totalt 142 ord (Scott, 2018). De filer som skapats av WordSmith Tools var sedan möjliga att analysera närmare i Microsoft Excel. WordSmith Tools inrymmer dock ingen funktion för att se hur mycket en ordlista täcker av en textsamling, vilket har resulterat i att jag utfört denna täckningsuträkning gällande typer och tokens manuellt. I korthet har en textsamling, i det här fallet F-SCUSSE eller LOCNESS, jämförts med en ordlista, i form av endera AWL eller AVL genom att skapa en WordList. Filerna har sedan exporterats till Excel där de ord som förekommer i två av texterna har markerats.

I det här skedet bör man vara varsam, eftersom en av träffarna kommer att härstamma från de akademiska ordlistorna. Därför bör man subtrahera en träff från varje akademiskt ords träff. De här momenten möjliggör i sin tur en procentuträkning.

Resultat

Resultaten kan sägas bestå av två delar. I den första delen presenteras generella frekvenser av ord, det vill säga de mest frekventa orden i F-SCUSSE, LOCNESS, samt BNC, både naturliga och när en stopplista har använts för att filtrera ut funktionsorden. I den andra delen är fokus på de akademiska ordlistorna samt på hur mycket de här akademiska orden förekommer i de två inlärar-korpusarna. Här presenteras täckning, både vad gäller typ-täckning och token-täckning, samt de mest vanligt förekommande akademiska orden i vardera korpusarna.

F-SCUSSE, LOCNESS och BNC har många likheter när det gäller de 15 mest frekventa orden i varje korpus, men ordningsföljden skiljer sig något. I F-SCUSSE dyker dock ett unikt ord upp, nämligen det personliga pronomenet *you*. F-SCUSSE och LOCNESS har till och med större likheter med varandra än vad LOCNESS och BNC har. När stopplistan används ignoreras de mest vanliga funktionsorden och enbart innehållsorden kvarstår. Samtidigt framträder somliga vaga ord (jfr Ringbom, 1998; 1999) i F-SCUSSE, till exempel *people*, *think*, *good*, *things* och *important*.

People var nästintill identiskt vad gäller frekvens i de två inlärarkorpusarna, och därför genomfördes en specifik konkordans-sökning på det här ordet. Det visade sig att användningen av ordet istället skiljer sig med avseende på det närmast framförliggande ordet. I F-SCUSSE föregås *people* oftast av *some*, *many* och *other*, medan ordet föregås i LOCNESS oftast av *many*, *the* och *more*. I LOCNESS förekom faktiskt ordkombinationen *other people* inte en endaste gång.

När en stopplista används, kan de mest frekventa orden i LOCNESS trots allt beskrivas som mycket mer ämnesspecifika än motsvarande lista för F-SCUSSE. Av de 20 mest frekventa orden i LOCNESS kan 11 stycken klassas som ämnesspecifika substantiv medan motsvarande siffra i F-SCUSSE är enbart ett ämnesspecifikt substantiv.

Finlandssvenska och engelska elever använder sig av en liknande akademisk vokabulär enligt AWL och AVL. Det här blir speciellt påtagligt med orden i AVL, till exempel *however*, *example*, *use*, *need* och *both*. Orden inkluderade i den

listan härstammar från många ordklasser, och de kan därför sägas ha många funktioner i en argumenterande text. Därför är det särskilt intressant att just sådana ord förekommer enligt liknande mönster bland båda grupper av elever.

Det var också möjligt att studera de ordfamiljer från AWL och AVL som inte var representerade i F-SCUSSE. De flesta ordfamiljerna återfanns dock i LOCNESS (i 21 av 24 fall). Allt det här tyder alltså på att modersmålstalarna använder sig av en mer teknisk vokabulär, men framförallt att de besitter en rikare vokabulär.

Sammanfattande diskussion

Resultaten visar att de båda akademiska ordlistorna täcker en högre andel ord i texter skrivna av elever med engelska som förstaspråk. Skillnaden gentemot de med engelska som främmande språk är dock inte så starkt framträdande som man kunde ha trott. Speciellt finlandssvenska elever som befinner sig på en högre nivå tycks använda sig av en vokabulär som påminner avsevärt om modersmålstalarnas. Av resultaten kan vi också utläsa att elever som studerar engelska som främmande språk och elever med engelska som förstaspråk uppvisar liknande tendenser i sitt språkbruk. Det här skulle antyda att elever som befinner sig i ungefär samma livssituation uttrycker sig på ett liknande vis, oavsett deras förstaspråk.

Jag har även uppmärksammat fenomen utöver de som anknyter till mina forskningsfrågor, vilket är vanligt inom LCR (Hasselgård, 2009: 238). Vaga ord (jfr Hasselgren, 1994; Leech, 2011) förekommer i både finlandssvenska och engelska elevers texter, och det kan vara nödvändigt att se på ett ord i dess samman-hang för att finna skillnader mellan de två grupperna, som i exemplet *people*.

Appendices

Appendix A. Document distributed to English teachers

Åbo Akademi

André Sandberg, fil. kand.
FHPT, Engelska språket och Litteraturen
+358 503078188
andre.sandberg@abo.fi
Handledare: Dr. Brita Wårvik
brita.warvik@abo.fi

Forskningsplan

Vocabulary Frequencies among Finland Swedish Upper Secondary School Students
(preliminär titel)

Ändamål
Ändamålet med den här studien är att undersöka finlandssvenska gymnasiestuderandes vokabulär i fritt formulerade texter. Genom att samla in genuina texter skrivna av gymnasiestuderande är målet att bygga upp en korpus i mindre format, bestående av till exempel 100-150 texter. Den här korpusen skulle sedan bilda material för min avhandling pro gradu.

Texterna
Lärarna skulle skicka texterna anonymt till mig, och de skulle även förbli anonyma. Skolorna skulle även de förbli onämnda. De enda som skulle ha tillgång till materialet skulle vara jag samt min handledare. Materialet skulle användas enbart för det här forskningsändamålet.

Analysen
Med hjälp av korpusverktyg som konkordansprogram kommer sedan materialet att bearbetas. Jag är intresserad av frekvenser av olika ord i som skulle förekomma i de olika texterna, om det exempelvis förekommer ord som är överanvända eller underanvända. Jag måste dock poängtera att jag inte är intresserad av vilka enskilda fel eleverna gör. I stället handlar om en kartläggning av elevernas produktiva vokabulär i skrift, det vill säga förekommer det mönster eller tendenser hos finlandssvenska gymnasiestuderandes engelska.

Resultat
Resultaten kommer, till den mån det är möjligt, att jämföras med tidigare studier, såsom studier baserade på ICLE (texter skrivna på engelska av finlandssvenska universitetsstuderande). På så sätt kan man förhoppningsvis även se kronologiska skillnader. Resultaten skulle även ge värdefull information om hur finlandssvenska gymnasiestuderanden skriver på engelska.

Appendix B. Document distributed to parents of minors.

Åbo Akademi

Åbo 21.5.2018

Bästa vårdnadshavare till elev i (skolans namn)

Jag studerar engelska på lärarlinjen vid Åbo Akademi och skriver för tillfället min avhandling pro gradu om finlandssvenska gymnasiestuderandens vokabulärkunskaper. Inom ramen för projektet samlar jag just nu in material för att bygga upp en korpus, det vill säga en samling språklig data, bestående av riktiga texter skrivna av gymnasiestuderande. Jag är intresserad av frekvenser av ord och vilka mönster som framkommer ur texterna. (obs. inte skrivfel)

I praktiken innebär det att ert barns engelskalärare skickar anonyma textfiler som eleven skrivit till mig och dessa filer buntas sedan samman med textfiler skrivna av elever från andra finlandssvenska gymnasier.

Resultaten behandlas konfidentiellt och används endast för forskningsändamål. Anonymitet garanteras och inte heller skolans namn kommer att framkomma.

Ert barns text är väldigt viktig för forskningsuppgiften. Vänligen underteckna blanketten om ni samtycker.

Vid frågor kontakta André Sandberg per e-post.

Jag _____
(vårdnadshavarens namn)

☐ godkänner

☐ godkänner inte

att _____s engelskatext får skickas
(elevens namn)

anonymt till André Sandberg för forskningsändamålet.

André Sandberg, fil. kand.
Engelskstuderande
Engelska språket och litteraturen
Åbo Akademi
Fabriksgatan 2, 20500 Åbo
E-post: andre.sandberg@abo.fi

References

Primary Sources

Coxhead, A. 2000a. *The Academic Word List*. Available: https://www.victoria.ac.nz/lals/resources/academicwordlist [07.02.2018]

Gardner D. & Davies, M. 2013a. *The Academic Vocabulary List*. Available: https://www.academicvocabulary.info/ [21.02.2018]

The British National Corpus (BNC), version 3 (BNC XML Edition). 2007. Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. Available: http://www.natcorp.ox.ac.uk/ [07.02.2018]

The Louvain Corpus of Native English Essays (LOCNESS). Created by the Centre for English Corpus Linguistics (CECL), Université catholique de Louvain, Belgium. Available: http://www.learnercorpusassociation.org/ [07.02.2018]

Academic Writing Sample

Callies, M. 2015. "Using Learner Corpora in Language Testing and Assessment: Current Practice and Future Challenges". In Castello, E. et al. 2015. *Studies in Learner Corpus Linguistics*. Peter Lang. Available: ProQuest Ebook Central [23.10.2017]

Gilquin, S. 2008. "What You Think Ain't What You Get: Highly polysemous verbs in mind and language". In Lapaire, J. et al. (eds.). 2008. *Du fait grammatical au fait cognitif. From Gram to Mind: Grammar as Cognition. Volume 2*. 235-55. Presses Universitaires de Bordeaux: Pessac.

Hulstjin, J. H. & Laufer, B. 2001. "Some empirical evidence for the involvement load hypothesis in vocabulary acquisition". In *Language Learning*. September, 2001. 51(3). 539-58.

Iwahori, Y. 2008. "Developing fluency: A study of extensive reading in EFL". In *Reading in a Foreign Language*. April, 2008. 20(1). 70-91.

Secondary Sources

Ai, H. & Lu, X. 2013. "A corpus-based comparison of syntactic complexity in NNS and NS university students' writing". In Diaz-Negrillo, A. et al. (eds.). 2013. *Automatic Treatment and Analysis of Learner Corpus Data*. 249-64. Amsterdam & Philadelphia: John Benjamins.

Altenberg, B. 2011. "Preface". In Meunier, F. et al. (eds.). 2011. *Taste of Corpora: In Honour of Sylviane Granger*. xiii-xv. Amsterdam & Philadelphia: John Benjamins.

BNC. 2015. "What is the BNC?". Available: http://www.natcorp.ox.ac.uk/corpus/index.xml [10.09.2018]

Browne, C., Culligan, B., & Phillips, J. 2013. *The New Academic Word List*. Available: http://www.newacademicwordlist.org/ [18.11.2018]

Callies, M. 2015. "Using Learner Corpora in Language Testing and Assessment: Current Practice and Future Challenges". In Castello, E. et al. 2015. *Studies in Learner Corpus Linguistics*. Peter Lang. Available: ProQuest Ebook Central [23.10.2017]

Callies, M. 2015a. "Learner corpus methodology". In Granger, S. et al. (eds.). 2015. *The Cambridge Handbook of Learner Corpus Research*. 35-55. Cambridge: Cambridge University.

Chambers, A. 2015. "The learner corpus as a pedagogic corpus". In Granger, S. et al. (eds.). 2015. *The Cambridge Handbook of Learner Corpus Research*. 445-64. Cambridge: Cambridge University.

Cheng, W., Warren, M & Xu, X. 2003. "The language learner as language researcher: Corpus linguistics on the timetable". In *System*. 2003: 31(2): 173-86.

Cobb, T. & Horst, M. 2015. "Learner corpora and lexis". In Granger, S. et al. (eds.). 2015. *The Cambridge Handbook of Learner Corpus Research*. 185-206. Cambridge: Cambridge University.

Cook, V. 1986. "The basis for an experimental approach to second language learning". In Cook, V. (ed.). 1986. *Experimental Approaches to Second Language Learning*. 3-21. Oxford: Pergamon.

Coxhead, A. 2000. "A new academic word list". In *TESOL Quarterly*. 2000: 34(2): 213-38.

Coxhead, A. 2011. "The academic word list 10 years on: Research and teaching implications". In *TESOL Quarterly*. 2011: 45(2): 355-61.

Davies, M. 2004. "Student use of large, annotated corpora to analyze syntactic variation". In Aston, G et al. 2004. *Corpora and Language Learners*. 259-69. Amsterdam & Philadelphia: John Benjamins.

De Cock, S & Granger, S. 2004. "Computer learner corpora and monolingual learners' dictionaries: The perfect match". In W. Teubert & M. Mahlberg (eds). 2004. *The Corpus Approach to Lexicography*. Special issue of *Lexicographica* 20: 72-86.

Doró, K. 2015. "Changes in the lexical measures of undergraduate EFL students' argumentative essays". In Pietilä, P. et al (eds.). 2015. *Lexical Issues in L2 Writing*. 57-76. Newcastle: Cambridge Scholars.

Erman, B. 2015. "Two different methodologies in the identification of recurrent word combinations in English L2 writing". In In Pietilä, P. et al (eds.). 2015. *Lexical Issues in L2 Writing*. 177-206. Newcastle: Cambridge Scholars.

Flowerdew, L. 2015. "Learner corpora and language for academic and specific purposes". In Granger, S. et al. (eds.). 2015. *The Cambridge Handbook of Learner Corpus Research*. 465-84. Cambridge: Cambridge University.

Fries, C. C. 1945. *Teaching and Learning English as a Foreign Language*. Ann Arbor: University of Michigan.

Gardner D. & Davies, M. 2013. "A New Academic Vocabulary List". In *Applied Linguistics*. 2014: 35(3): 305-327. Available: https://academic.oup.com/journals [07.02.2018]

Garside, R. & Smith, N. 1997. "A hybrid grammatical tagger: CLAWS4". In Garside, R. et al. (eds.). 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. 102-121. London: Longman.

Gilquin, G. 2015. "From design to collection of learner corpora". In Granger, S. et al. (eds.). 2015. *The Cambridge Handbook of Learner Corpus Research*. 9-34. Cambridge: Cambridge University.

Granger, S. 2002a. "A bird's eye view of learner corpus research". In Granger, S. et al. 2002. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. 3-33. Amsterdam & Philadelphia: John Benjamins. (

Granger, S. 2008. “Learner Corpora”. In Lüddeling, A. & Kytö, M (eds.). *Corpus Linguistics. An International Handbook: Volume 1*. 259-75. Berlin: Mouton de Gruyter.

Granger, S. 2009. “The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation”. In Aijmer, K. (ed.) *Corpora and Language Teaching*. 13-32. Amsterdam: John Benjamins.

Granger, S. et al. 2015. “Introduction: learner corpus research – past, present and future”. In Granger, S. et al. (eds.). 2015. *The Cambridge Handbook of Learner Corpus Research*. 1-5. Cambridge: Cambridge University.

Granger, S., Dagneaux, E. & Meunier, F. (eds.). 2002. *International Corpus of Learner English*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Granger, S. & Paquot, M. 2009. “In search of General Academic English: A corpus-driven study”. In Katsampoxaki-Hodgetts, K. (ed.) *Options and Practices of L.S.P practitioners Conference Proceedings*. 94-108. University of Crete Publications, E-media.

Granger, S. & Paquot, M. 2009a. ”Lexical verbs in academic discourse: a corpus-driven study of learner use”. In Charles, M., Pecorari, D. & Hunston, S. (eds.). 2009. *Academic Writing*. *At the Interface of Corpus and Discourse*. 193-214. London & New York: Continuum.

Granger, S. & Rayson, P. 1998. “Automatic lexical profiling of learner texts”. In Granger, S. (ed.). *Learner English on Computer*. 119-31. London: Longman.

Gries, St. Th. 2015. “Statistics for learner corpus research”. In Granger, S. et al. (eds.). 2015. *The Cambridge Handbook of Learner Corpus Research*. 159-81. Cambridge: Cambridge University.

Hartshorn, K. J. & Hart, J. M. 2016. “Comparing the academic word list with the academic vocabulary list: analyses of frequency and performance of English language learners”. The Journal of Language Teaching and Learning, 6(2), 70-87.

Hasselgren, A. n.d.. *EVA Corpus of Norwegian School English*. Information: http://clu.uni.no/icame/ij21/eva-corp.pdf [18.11.2018]

Hasselgren, A. 1994. “Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary”. In *ITL* 1994, 4(2). 237-60.

Hasselgård, H. 2009. “Thematic choice and expressions of stance in English argumentative texts by Norwegian learners”. In Aijmer, K. (ed.) *Corpora and Language Teaching*. 121-39. Amsterdam: John Benjamins.

Hasselgård, H. & Johansson, S. 2011. “Learner corpora and contrastive interlanguage analysis”. In Meunier, F. et al. (eds.). 2011. *Taste of Corpora: In Honour of Sylviane Granger*. 33-61. Amsterdam & Philadelphia: John Benjamins.

Hernandez, M. M. 2017. *Comparing the AWL and AVL in Textbooks from an Intensive English Program*. Master’s thesis. Provo: Brigham Young University.

Hinkel, E. 2003. *Teaching Academic ESL Writing: Practical Techniques in Vocabulary and Grammar*. New York & Abingdon: Routledge.

Jarvis, S. & Paquot, M. 2015. “Learner corpora and native language identification”. In Granger, S. et al. (eds.). 2015. *The Cambridge Handbook of Learner Corpus Research*. 606-27. Cambridge: Cambridge University.

Leech, G. 1998. “Preface”. In Granger, S. (ed.). *Learner English on Computer*. xiv-xx. London & New York: Longman.

Leech, G. 2011. “Frequency, corpora and language learning”. In Meunier, F. et al. (eds.). 2011. *Taste of Corpora: In Honour of Sylviane Granger*. 7-31. Amsterdam & Philadelphia: John Benjamins.

Leppänen, S. et al. 2011. *National Survey on the English Language in Finland: Uses, Meanings and Attitudes*. Available: http://www.helsinki.fi/varieng/series/volumes/05/index.html [08.03.2018]

Lindgrén, S. 2015. “Academic vocabulary and readability in EFL theses”. In Pietilä, P. et al (eds.) *Lexical Issues in L2 Writing*. 155-74. Newcastle: Cambridge Scholars.

Lozano, C. & Mendikoetxea, A. 2013. “Learner corpora and second language acquisition: The design and collection of CEDEL2”. In Diaz-Negrillo, A. et al. (eds.). 2013. *Automatic Treatment and Analysis of Learner Corpus Data*. 65-100. Amsterdam & Philadelphia: John Benjamins.

Mauranen, A. 2011. “Learners and users – Who do we want corpus data from?”. In Meunier, F. et al. (eds.). 2011. *Taste of Corpora: In Honour of Sylviane Granger*. 155-71. Amsterdam & Philadelphia: John Benjamins.

Meunier, F. 2015. “Developmental patterns in learner corpora”. In Granger, S. et al. (eds.). 2015. *The Cambridge Handbook of Learner Corpus Research*. 379-400. Cambridge: Cambridge University.

Ministry of Education and Culture, Finland. 2018. "Introduction of matriculation examination in English is progressing". Published 09.03.2018. Available: https://minedu.fi/en/article/-/asset_publisher/englanninkielinen-ylioppilastutkinto-etenee [18.11.2018]

Myles, F. 2015. "Second language acquisition and learner corpus research". In Granger, S. et al. (eds.). 2015. *The Cambridge Handbook of Learner Corpus Research*. 309-31. Cambridge: Cambridge University.

Möller, V. 2017. *Language Acquisition in CLIL and Non-CLIL Settings: Learner Corpus and Experimental Evidence on Passive Constructions.* Amsterdam & Philadelphia: John Benjamins.

Nation, I. S. P. 1990. *Teaching and Learning Vocabulary*. Boston: Heinle & Heinle.

Nation, I. S. P. 2001. *Learning Vocabulary in Another Language*. 7th printing. Cambridge: Cambridge University.

Nesselhauf, N. 2004. "How learner corpus analysis can contribute to language teaching: a study of support verb constructions". In Aston, G et al. 2004. *Corpora and Language Learners*. 109-24. Amsterdam & Philadelphia: John Benjamins.

Newman, J. A. 2017. *A Corpus-Based Comparison of the Academic Word List and the Academic Vocabulary List*. Master's thesis. Provo: Brigham Young University.

Olsson, E. & Sylvén, L. K. 2017. "Validity in high- and low-stakes tests: A comparison of academic vocabulary and some lexical features in CLIL and non-CLIL students' written texts". In *Academic Language in a Nordic Setting – Linguistic and Educational Perspectives* 2017: 9(3): 127-46. Available: https://www.journals.uio.no/ [29.01.2018]

OED. 2017. "Dictionary facts". Available: https://public.oed.com/history-of-the-oed/dictionary-facts/ [07.05.2018]

Paquot, M. 2010. *Academic Vocabulary in Learner Writing: From Extraction to Analysis.* London & New-York: Continuum.

Pípalová, R. 2015. "Reporting verbs in native and non-native academic discourse". In Pietilä, P. et al (eds.) *Lexical Issues in L2 Writing*. 127-54. Newcastle: Cambridge Scholars.

Prentice, J. 2017. "Infrastruktur för svensk andraspråksforskning (och annan svensk språkforskning: Möten mellan andraspråksforskning och datalingvistik".

Presentation at ATDS, Kiel. 27-29 Sep. 2017. Available: https://spraakbanken.gu.se/sites/spraakbanken.gu.se/files/2017_sept_Plenar_Kiel_Julia.pdf [06.06.2018]

Pursiainen, J. et al. 2017. "Lukion ainevalinnat ja opiskelijarekrytointi" Part of the *AVAIN* project. University of Oulu. Available: http://www.oulu.fi/avain/ [10.04.2018]

Rayson, P. & Baron, A. 2011. "Automatic error tagging of spelling mistakes in learner corpora". In Meunier, F. et al. (eds.). 2011. *Taste of Corpora: In Honour of Sylviane Granger*. 109-26. Amsterdam & Philadelphia: John Benjamins.

Ringbom, H. 1998. "Vocabulary frequencies in advanced learner English: A cross-linguistic approach". In Granger S. (ed.). 1998. *Learner English on Computer*. 41-52. London & New York: Addison Wesley Longman.

Ringbom, H. 1999. "High-frequency verbs in the ICLE corpus". In Renouf A. (ed.) *Explorations in Corpus Linguistics*. 191-200. Amsterdam and Atlanta: Rodopi.

Salazar, D. 2014. *Lexical Bundles in Native and Non-Native Writing: Applying a corpus-based study to language teaching*. Amsterdam & Philadelphia: John Benjamins.

Schmitt, D & Schmitt, N. 2005. *Focus on Vocabulary: Mastering the Academic Word List*. White Plains, NY: Longman Pearson.

Schmitt, N. 2010. *Researching Vocabulary: A Vocabulary Research Manual*. Houndmills: Palgrave Macmillan.

Schmitt, N. & Redwood, S. 2011. "Learner knowledge of phrasal verbs: A corpus-informed study". In Meunier, F. et al. (eds.). 2011. *Taste of Corpora: In Honour of Sylviane Granger*. 173-208. Amsterdam & Philadelphia: John Benjamins.

Scott, M. 2018. WordSmith Tools version 7. Stroud: Lexical Analysis Software.

Scott, M. 2018a. WordSmith Tools Help. Stroud: Lexical Analysis Software.

Scott, M & Tribble, C. 2006. *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam & Philadelphia: John Benjamins.

Seidlhofer, B. 2002. "*Habeas corpus* and *divide et impera*: 'Global English' and applied linguistics". In Spelman Miller, K. & Thompson, P. (eds). *Unity and Diversity in Language Use*. 198-220. London: Continuum.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. 2004. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

Studentexamensnämnden's. homepage. 2018 Available: https://www.ylioppilastutkinto.fi/sv/ [30.01.2018]

Tono, Y. et al. 2012. *Developmental and Crosslinguistic Perspectives in Learner Corpus Research*. Amsterdam & Philadelphia: John Benjamins.

Utbildningsstyrelsen (Finnish National Agency for Education). 2014. Educational Statistics' Yearbook. Available: http://www.oph.fi/publikationer/2014/arsbok_for_utbildningsstatistik_2014 [08.02.2018]

Viana, V. 2007. "Corpus linguistics, language learning & ELT: interviewing Sylviane Granger". APLIERJ Newsletter. 2007(1): 11-14.

Virtanen, T. 1998. "Direct questions in argumentative student writing". In Granger S. (ed.) 1998. *Learner English on Computer*. 94-106. London & New York: Addison Wesley Longman.

West, M. 1953. *A General Service List of English Words*. London: Longman.

Xue, G. & Nation, I. S. P. 1984. "A university word list". In *Language Learning and Communication*. 3: 215-229.

Zwiers, J. 2008. *Building Academic Language: Essential Practices for Content Classrooms.* Hoboken: John Wiley & Sons.

Ädel, A. 2010. "Using corpora to teach academic writing: Challenges for the direct approach". In Campoy-Cubillo, M. C. et al. 2010. *Corpus-Based Approaches to English Language Teaching*. 39-54. London & New York: Continuum.

Ädel, A. 2015. "Variability in learner corpora". In Granger, S. et al. (eds.). 2015. *The Cambridge Handbook of Learner Corpus Research*. 401-21. Cambridge: Cambridge University.